Hot-Deck-Verfahren zur Imputation fehlender Daten

—

Auswirkungen des Donor-Limits

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum politicarum
an der Fakultät für Wirtschaftswissenschaften und Medien
der Technischen Universität Ilmenau

Vorgelegt von:
Dipl. Wirtsch.-Ing.
Dieter William Hermann Joenssen
aus Düsseldorf

Gutachter:
Univ.-Prof. Dr. rer. pol. habil. Udo Bankhofer
Univ.-Prof. Dr. rer. pol. habil. Ralf Trost

Abgabedatum:
2014/11/24
Disputationsdatum:
2015/05/12

urn:nbn:de:gbv:ilm1-2014000595

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung und Zielsetzung

Im Idealfall sind alle Datensätze vollständig. Somit sollte die Entstehung von fehlenden Daten grundsätzlich verhindert werden, denn die Gründe für Datenausfall sind in vielen Fällen auf fehler- und mangelhafte Untersuchungsdesigns zurückzuführen. Daher ist es auch nachvollziehbar, dass ein Auftreten von fehlenden Werten unzufriedenstellend ist. Jedoch lässt sich nicht jede Untersuchung mit vorhandenen ökonomischen Beschränkungen so konzeptionieren und durchführen, dass der Datenausfall grundsätzlich verhindert wird. Zeitliche und personelle Beschränkungen können unterbinden, dass gewisse Werte für alle Untersuchungssubjekte gemessen werden, da gewisse Subpopulationen nur schwer zu erreichen sind. Nacherhebungen führen bei fehlenden Angaben nicht unbedingt zu einer Beantwortung der offenen Fragen. Veränderungen in den Lebenssituationen von Personen verhindern, dass diese für die gesamte Dauer einer Studie zur Verfügung stehen. Zudem können bei technischen Systemen notwendige Redundanzen in der Sensorik aus physikalischen Gründen nicht immer realisiert werden.

Eine Vorgehensweise, die das Entstehen von fehlenden Daten in Kauf nimmt, ist somit einer, die dieses stets zu verhindern versucht, in gewissen Situationen nicht nur vorzuziehen, sondern auch unumgänglich. Die Forschung nach besseren Methoden zum Umgang mit fehlenden Werten wird daher durch diese Erkenntnis angetrieben. Zwar hat das Forschungsfeld in den letzten Jahren an Beachtung und Anerkennung gewonnen, jedoch wurde dessen Relevanz noch nicht vollständig von der empirischen Forschung und Praxis angenommen.

Mit unvollständigen Daten korrekt zu verfahren ist wichtig, da der Um-

gang mit dem Datenausfall immer vor jeglichen Analysen und Modellbildungen steht. Hierdurch wird klar, dass ein inkorrekter Umgang mit den fehlenden Werten zu schlechteren Analysen und Modellen führt, was wiederum in schlechtere oder gar falsche Entscheidungen münden kann. Für keinen Bereich, in dem quantitative Methoden verwendet werden, scheint dies wichtiger zu sein als in der Pharmakologie. Hier wurde im Jahr 1999 bei der „Europäischen Arzneimittel-Agentur“ (European Medicines Agency, EMA) der „Ausschuss für Humanarzneimittel“ (Committee for Medicinal Products for Human Use, CHMP) damit beauftragt, Richtlinien für den Umgang mit fehlenden Werten bei konfirmatorischen klinischen Studien zu entwerfen. Diese Richtlinien, die sich unmittelbar auf das Zulassungsverfahren von Arzneimitteln in der Europäischen Union auswirken, sind 2011 in Kraft getreten (Committee for Medicinal Products for Human Use, 2010). Ähnliche Entwicklungen sind auf der anderen Seite des Atlantiks zu beobachten. Zwar existieren in den USA noch keine verbindlichen Richtlinien zum Umgang mit fehlenden Daten (vgl. O'Kelly und Ratitch, 2014, S. 55), dennoch deuten aktuelle Entwicklungen darauf hin, dass diese in Zukunft zu erwarten sind. Das amerikanische Pendant zu der EMA, die Food and Drug Administration (FDA), beauftragte im Jahr 2008 die National Academy of Sciences mit der Erstellung von Empfehlungen zum Umgang mit fehlenden Daten in klinischen Studien. Diese Empfehlungen (Panel on Handling Missing Data in Clinical Trials, 2010) wurden bereits 2010 fertig gestellt und veröffentlicht. In beiden Werken wird sich eindeutig dazu bekannt, dass fehlende Daten weder bei der Planung noch der Durchführung und Interpretation der Analysen ignoriert werden dürfen (vgl. Committee for Medicinal Products for Human Use, 2010, S. 1 beziehungsweise Panel on Handling Missing Data in Clinical Trials, 2010, S. 55 ff.).

Aber auch in anderen Bereichen muss sich zunehmend mit den Auswirkungen von fehlenden Daten auseinandergesetzt werden, da die Methoden zum Umgang mit fehlenden Werten nicht nur Einfluss auf Entscheidungen, sondern auch auf Entscheidungsträger selber entfalten können. So musste sich beispielsweise der oberste Gerichtshof der Vereinigten Staaten 2001 damit beschäftigen, ob die Verwendung einer konkreten Methode zum Umgang mit fehlenden Daten verfassungsgemäß sei[1]. Hier hatte die Anwendung von Hot-Deck-Verfahren

[1] Siehe hierzu: Utah v. Evans, 536 U.S. 452 (2002).

durch das U.S. Census Bureau zu einer Erhöhung der Gesamtbevölkerung um ca. 1,2 Millionen Menschen geführt (vgl. Cantwell et al., 2004, S. 208). Durch die Ersetzung fehlender durch plausible Daten (Imputation) erhöhten sich die Bevölkerungszahlen der einzelnen Bundesstaaten ungleichmäßig. Beispielsweise stieg die Bevölkerung von North Carolina um 0,4% und die von Utah um 0,2%. Diese ungleichmäßige Erhöhung durch die Imputation mittels Hot-Deck-Verfahren führte dazu, dass North Carolina einen weiteren Sitz im Repräsentantenhaus erhielt; ein Sitz, der ohne die Auswirkungen der Imputation an den Bundesstaat Utah gegangen wäre (vgl. Barak und Fried, 2002, S. 201).

Trotz dieses Gewinns an Beachtung und Anerkennung sowie der Tatsache, dass zu den Methoden zum Umgang mit fehlenden Daten seit über 80 Jahren geforscht wird[2], finden fehlende Daten und der Umgang mit diesen unzureichend Erwähnung in quantitativen empirischen Untersuchungen. Besonders gut zu beobachten ist dies in wissenschaftlichen Untersuchungen im Bereich der Psychologie. Hier befasste sich bereits Ende der 1990er Jahre die American Psychological Association (APA) mit der Entwicklung von Richtlinien für den Umgang mit fehlenden Daten. In ersten Empfehlungen der eingesetzten Kommission wird explizit von Methoden abgeraten, die Fälle mit fehlenden Beobachtungen von Analysen ausschließen (vgl. Wilkinson und Task Force on Statistical Inference, 1999, S. 598). Derweil ist in den konkreten Richtlinien der APA, basierend auf den Empfehlungen der Kommission, festgehalten worden, dass jegliches Vorhandensein von fehlenden Daten sowie deren Umfang und Auswirkungen auf die Analysen zu dokumentieren sind (vgl. American Psychological Association, 2001). Jedoch werden diese konkreten Empfehlungen und Richtlinien kaum umgesetzt, wie die Untersuchungen von Peughd und Enders (2004) sowie Bodner (2006) zeigen. In ihrer Untersuchung von Studien aus dem Jahr 2003 stellen Peughd und Enders (2004, S. 541) fest, dass in 31% der Studien zwar fehlende Daten explizit diskutiert werden, jedoch nur in etwa 4% der Studien Methoden verwendet werden, von denen die APA nicht abrät (vgl. Peughd und Enders, 2004, S. 541 f.). Ähnliches wird auch in der Studie von Bodner (2006) vermerkt. Er stellt fest, dass nur in 36 von 181 untersuchten Studien explizit auf fehlende Werte eingegangen wird. Des Weiteren

[2] Beispielsweise beschäftigte sich bereits Wilks (1932) mit der Maximum-Likelihood-Schätzung der Parameter von unvollständig beobachteten, bivariat normalverteilten Merkmalen.

wird kommentiert, dass, sofern die Methoden zum Umgang mit den fehlenden Werten diskutiert wurden, in allen bis auf zwei Studien gerade jene Methoden verwendet wurden, von denen die APA abrät (Bodner, 2006, S. 677). Dieser nachlässige Umgang mit fehlenden Werten entgegen konkreter Empfehlungen lässt vermuten, dass in anderen Bereichen der Wissenschaft und der Praxis nicht weniger nachlässig verfahren wird.

Die Nachlässigkeit, welche noch im Umgang mit fehlenden Daten zu beobachten ist, kann sicherlich auch in Teilen auf eine unzureichende Erforschung der Methoden zum Umgang mit fehlenden Daten zurückgeführt werden. Es existiert weitestgehend Konsens in der Wissenschaft, dass keine einzelne Methode unter jeden Umständen die beste zur Behandlung von fehlenden Daten ist (vgl. Wilkinson und Task Force on Statistical Inference, 1999, S. 594; Committee for Medicinal Products for Human Use, 2010, S. 3, 5; Panel on Handling Missing Data in Clinical Trials, 2010, S. 2, 4). Dennoch bleibt nicht nur eine umfassende Rangordnung der Methoden in Abhängigkeit von der Anwendungssituation aus, sondern es fehlt bereits an Untersuchungen, die die Sinnhaftigkeit gewisser Verfahrensvarianten erkunden. Dieser Mangel an Forschung verunsichert Nutzer, die sich einer ständig wachsenden Vielzahl von Methoden und Verfahrensvarianten gegenübergestellt sehen, und dient als Motivation der empirischen Untersuchungen dieser Arbeit.

Den Untersuchungsgegenstand stellen die Hot-Deck-Verfahren, die ursprünglich in dem Census Bureau der Vereinigten Staaten erfunden wurden und auch Gegenstand des Gerichtsverfahrens von Utah v. Evans waren. Diese Verfahren wurden seit der Konzeption des ersten Urtyps in den 1930er Jahren stets weiterentwickelt, wobei die theoretischen Eigenschaften der entstandenen Varianten an Hot-Deck-Verfahren bis heute schlecht ergründet sind. Eine dieser Variationen stellt die Verwendung eines sogenannten Donor-Limits dar. Das Donor-Limit fand seit seiner ersten Erwähnung in der einschlägigen Literatur[3] zwar rege Anwendung, aber Auswirkungen dessen wurden bis dato noch nie untersucht (vgl. Andridge und Little, 2010, S. 43). Das erklärte Ziel dieser Arbeit besteht nun darin zu ergründen, ob und unter welchen Bedingungen die Verwendung eines Donor-Limits Vorteile bietet.

[3] Zum ersten Mal wird das Donor-Limit wohl von Sande (1983, S. 345) erwähnt, doch die Idee geht auf die Arbeit von Kalton und Kish (1981) zurück.

Kapitel 2 widmet sich zunächst den Grundlagen der Literatur zu fehlenden Daten. Behandelt werden folgende Fragen der Missing-Data-Problematik:

1. Weshalb fehlen die Daten?
2. Wo fehlen die Daten?
3. Wie fehlen die Daten?
4. Wie kann mit den fehlenden Daten umgegangen werden?

Die erste Frage zielt auf die vorhandene Ausfallursache ab und inwiefern Kausalzusammenhänge für den Ausfall konkreter Werte verantwortlich sind. Die zweite Frage weist auf die archetypischen Muster hin, die durch einen Datenausfall in einer sortierten Datenmatrix entstehen können. Wie die Daten fehlen, ist die Frage nach den grundlegenden Eigenschaften von jenem stochastischen Prozess, mittels dem der Datenausfall modelliert werden kann. Diese Frage ist zentral, sofern die Datenerhebung bereits stattgefunden hat. Die Antwort bestimmt, welche Methode zum Umgang mit fehlenden Daten angewendet werden kann oder ob eine sinnvolle Analyse der Datenbasis überhaupt möglich ist. Zwei Kategorien von Methoden, mittels denen fehlende Daten grundsätzlich behandelt werden können, werden anschließend dargestellt. Unterschieden wird zwischen Eliminierungs- und Imputationsverfahren. Hierbei erzeugen die Eliminierungsverfahren durch eine systematische Löschung von Untersuchungsobjekten oder deren Merkmalen eine vollständige Datenmatrix, welche mit herkömmlichen statistischen Methoden analysiert werden kann. Imputationsverfahren erstellen hingegen eine vollständige Datenmatrix durch die Ersetzung von fehlenden Werten durch geeignete Schätzungen.

Im **Kapitel 3** erfolgt eine nähere Darstellung der Hot-Deck-Verfahren. Hier werden als erstes die historische Entwicklung und die Herkunft dieser Verfahren beschrieben. Aufgrund konkurrierender Verständnisse bezüglich des Begriffs „Hot-Deck“ wird mittels einer umfangreichen Literaturanalyse eine für diese Arbeit notwendige Definition erarbeitet. Als nächstes erfolgt eine Darstellung der Verfahrensvarianten, die für Hot-Deck-Methoden möglich sind und in der Literatur diskutiert werden. Das Kapitel schließt mit einer Darstellung von Hot-Deck-Methoden als ganzzahliges Optimierungsproblem ab. Verbesserungen der Imputationsqualität, die sich durch eine Lösung des Optimierungspro-

blems ergeben können, werden erörtert und mit Hilfe einer Simulationsstudie abgeschätzt.

In **Kapitel 4** erfolgt schließlich mittels einer Voruntersuchung und einer konfirmatorischen Studie eine Betrachtung der Auswirkungen eines Donor-Limits. Beantwortet werden mit Hilfe zweier umfassender Simulationsstudien die folgenden Fragen:

1. Ist grundsätzlich ein Donor-Limit im Rahmen einer Hot-Deck-Imputation sinnvoll beziehungsweise notwendig?

2. Von welchen Gegebenheiten des vorliegenden Datenmaterials hängt eine notwendige Beschränkung ab?

3. Ist die Vorteilhaftigkeit eines Donor-Limits abhängig vom verwendeten Hot-Deck-Verfahren?

4. Können in den jeweils betrachteten Fällen weitere Empfehlungen hinsichtlich des Donor-Limits gegeben werden?

Zur Unterstützung der Erläuterung der vorgestellten Konzepte und Methoden werden in den Kapiteln 2 und 3 Beispiele eingesetzt. Diese, von dem restlichen Text durch den Hinweis „**Beispiel**“ getrennten Abschnitte, verwenden, mit einigen Ausnahmen, entweder simulierte Daten oder jene, die im Anhang A tabellarisch dargestellt sind. Simulationen wurden mittels der Statistiksoftware R in der Version 3.0.2 (R Core Team, 2013) mit dem (Pseudo-) Zufallszahlengenerator „Mersenne-Twister“ (Matsumoto und Nishimura, 1998) durchgeführt. Multidimensional normalverteilte Zufallszahlen wurden mit der Funktion *rmvnorm* aus dem R-Paket **mvtnorm** (Genz et al., 2013) generiert. Bei den Daten, die im Anhang A tabellarisch dargestellt sind, handelt es sich um eine einfache Stichprobe von elf Merkmalen der American Community Survey (ACS) aus 2009 vom 28. September 2010 (Ruggles et al., 2010; U.S. Bureau of the Census, 2010), welche vollständig unter https://usa.ipums.org/usa/ verfügbar sind.

Kapitel 2

Die Missing-Data-Problematik

Die Beschreibung der Missing-Data-Problematik geht in der Literatur mit unterschiedlichen Bezeichnungen einher. Schlagworte wie „missing-“, „partial-“, „incomplete-“, „fragementary-“, „omitted-“ und „spoilt-data“ sowie „missing-values“ und „-units“, „subject-“ und „item-nonresponse“, „unobserved-units“ und „answer-refusal“ stellen nur einen Auszug der Begrifflichkeiten dar, unter denen in der englischsprachigen Literatur fehlende Daten und Probleme, die mit diesen einhergehen, diskutiert werden. Anhand dieser sehr heterogenen Liste wird deutlich, dass fehlende Daten in den unterschiedlichsten Bereichen bereits thematisiert wurden. Enders (2010, S. 1) bezeichnet fehlende Daten in der Sozial-, Verhaltens- und Medizinforschung sogar als allgegenwärtig. Zwar sind fehlende Daten nicht die einzigen Ursachen von Verzerrungen in solchen Erhebungen, aber wohl die auffälligsten. Daher gewinnt jene Forschung, die sich mit fehlenden Daten auseinandersetzt, an Bedeutung. Dies resultiert nicht zuletzt auch daraus, dass immer mehr Forschungsbereiche, wie bereits in Kapitel 1 dargelegt, einen fundierten Umgang mit fehlenden Daten fordern. Ein solcher Umgang kann mittels einer Auseinandersetzung mit den folgenden vier zentralen Fragen der Missing-Data-Problematik erreicht werden:

1. Was ist für den Ausfall der Daten kausal?
2. Wo treten die fehlenden Werte innerhalb der Datenmatrix auf?
3. Wie lässt sich der Datenausfall als Zufallsprozess modellieren?
4. Welche Methoden lassen sich zum Umgang mit einer unvollständigen Datenmatrix anwenden?

Die erste Frage lässt sich mittels der Ausführungen in Abschnitt 2.1 zu den Ausfallursachen beantworten. Archetypische Muster fehlender Daten werden zur Beantwortung der zweiten Frage in Abschnitt 2.2 dargestellt. Die für den Umgang mit fehlenden Daten unumgängliche Frage nach der Art des vorhandenen Ausfallmechanismus wird in dem vorletzten Abschnitt 2.3 erörtert. Der letzte Abschnitt widmet sich der Darstellung eines Auszugs an Missing-Data-Methoden, mit deren Hilfe sich fehlende Daten behandeln lassen.

Eine sinnvolle Erörterung der Theorien und Grundlagen der folgenden Abschnitte erfordert die Festlegung einer einheitlichen Notation. Grundsätzlich ist der Ausgangspunkt nachfolgender Betrachtungen immer eine unvollständige Datenmatrix A der Form

$$A = (a_{ik})_{n,m} = \begin{pmatrix} a_{11} & \dots & \dots & a_{1m} \\ \vdots & \circ & & \vdots \\ & & & \circ \\ \vdots & & \circ & \vdots \\ a_{n1} & \dots & \dots & a_{nm} \end{pmatrix}, \tag{2.1}$$

wobei das Symbol $\circ$ fehlende Ausprägungen andeutet. n bezeichnet die Anzahl an Objekten, deren Untersuchung grundsätzlich von Interesse ist und mittels der m Merkmale charakterisiert werden. Des Weiteren ist es zur Darstellung verschiedener Konzepte nützlich, eine Missing-Data-Indikatormatrix (MD-Indikatormatrix) V der Form

$$V = (v_{ik})_{n,m} = \begin{pmatrix} v_{11} & \dots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{nm} \end{pmatrix} \text{ mit } v_{ik} = \begin{cases} 1 & \text{falls } a_{ik} \text{ fehlt} \\ 0 & \text{sonst} \end{cases} \tag{2.2}$$

zu definieren. Die Summierung einer Spalte k von V resultiert in der Anzahl der fehlenden Werte $v_{\bullet k}^{mis}$ für dieses k-te Merkmal. Der Anteil der fehlenden Werte in einem Merkmal k wird mit $\tilde{v}_{\bullet k}^{mis}$ bezeichnet. Die Symbolik für eine objektweise Betrachtung erfolgt mit $v_{i\bullet}^{mis}$ beziehungsweise $\tilde{v}_{i\bullet}^{mis}$ für das i-te Objekt analog. Bezogen auf die gesamte Datenmatrix wird v^{mis} für die Anzahl und $\tilde{v}^{mis}$ für den Anteil von fehlenden Werten geschrieben. Die Anzahl der Merkmale beziehungsweise Objekte, die mindestens einen fehlenden Wert aufweisen, beträgt q beziehungsweise r.

Die tatsächlichen, nicht bekannten, fehlenden Werte der Datenmatrix A werden in A^{mis} und die vorhandenen Werte in A^{obs} zusammengefasst, so dass

$$A^{mis} := \{a_{ik} | \forall i \in N, k \in M : v_{ik} = 1\} \tag{2.3}$$

und

$$A^{obs} := \{a_{ik} | \forall i \in N, k \in M : v_{ik} = 0\}, \tag{2.4}$$

wobei $N = \{1, \ldots, n\}$ beziehungsweise $M = \{1, \ldots, m\}$ der Menge aller betrachteten Objekte beziehungsweise Merkmale entspricht. A^{mis} und A^{obs} stellen somit keine Matrizen dar (vgl. Abbildung 2.1), sondern werden als Kurzform für Erklärungszwecke verwendet (vgl. Bankhofer, 1995, S. 6 oder Little und Rubin, 2002, S. 11 f.).

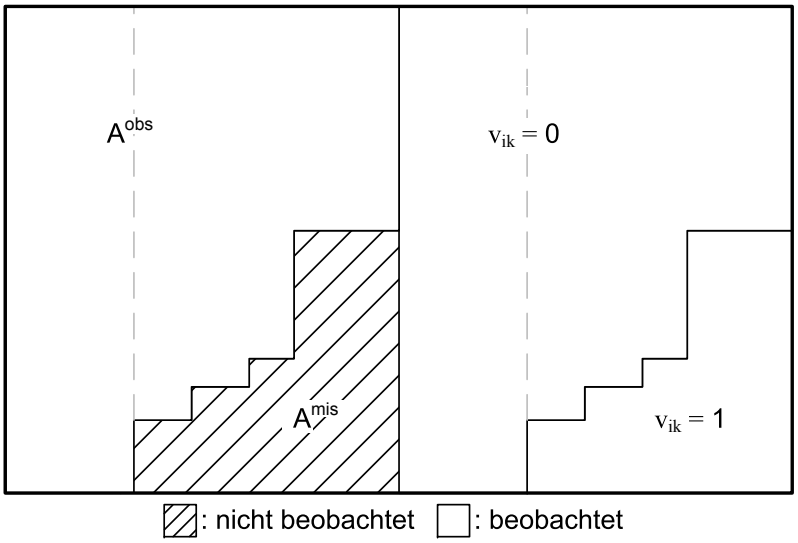


Abbildung 2.1: Visualisierung einer möglichen Aufteilung von A mit zugehörigem V (in Anlehnung an Little, 1982, S. 239)

2.1 Ausfallursachen

Untrennbar verbunden mit dem Auftreten fehlender Werte ist die Frage nach dem Grund für diesen Ausfall. Ausfallursachen für fehlende Daten sind Kausalzusammenhänge, die unabhängig von dem zugrundeliegenden stochastischen Prozess, der für den Ausfall konkreter Werte verantwortlich ist, sind. Bedeutend ist es, kausale Ursachen für den Ausfall von Werten zu identifizieren. Dies ist nicht nur wichtig, um ein allgemeines Verständnis für die vorliegenden

Daten zu erlangen, sondern insbesondere, um die vorliegende Art des Ausfallmechanismus im Anschluss erkennen zu können.

Bankhofer (1995, S. 8 ff.) gliedert mögliche Ausfallursachen in Anlehnung an Schnell (1986, S. 24–56) danach, in welcher der fünf Stufen des datenanalytischen Untersuchungsprozesses diese Ursachen auftreten. Gemäß Bankhofer (1995, S. 5) sind diese fünf Stufen:

- Festlegung von Untersuchungsgegenstand und Untersuchungsziel
- Diskussion der Datenbasis
- Datenerhebung
- Datenauswertung
- Interpretation der Ergebnisse

Von stärkerer Bedeutung für die Erkennung von Ausfallursachen, als der Zeitpunkt des Auftretens von fehlenden Werten im datenanalytischen Untersuchungsprozess, ist die Frage danach, wie die zweite und dritte Phase ausgestaltet sind. Hierfür ist es naheliegend, die Generierung der Datenbasis als eine Art Produktionssystem aufzufassen. Dieses Produktionssystem kann als soziotechisches System (Trist und Bamford, 1951) modelliert werden, welches aus einem gegebenen Untersuchungsdesign eine Datenbasis „produziert“.

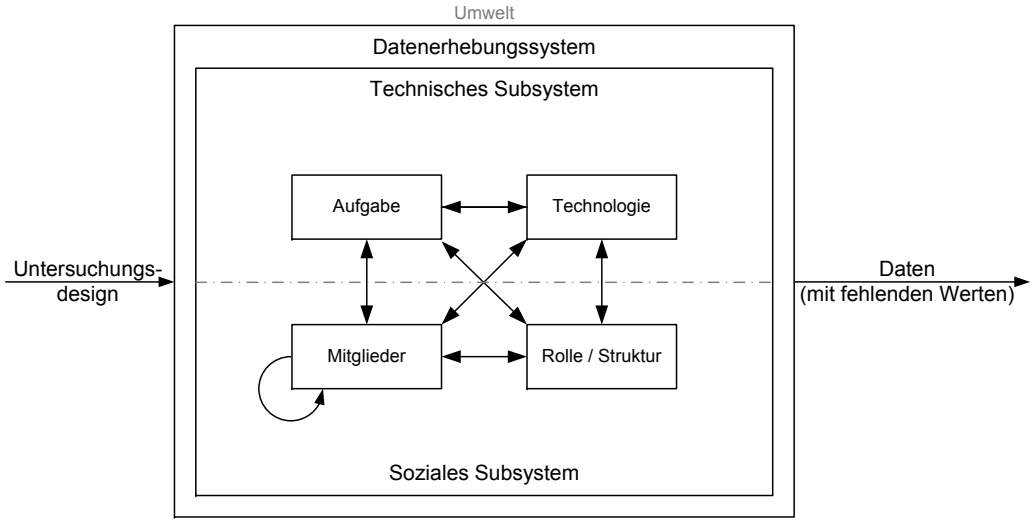


Abbildung 2.2: Das soziotechnische System in der Datenerhebung (in Anlehnung an Sydow, 1985, S. 29)

Die Abbildung 2.2 verbildlicht die Komponenten des soziotechnischen Systems in der Datenerhebung. An der Erstellung der Daten aus einem gegebenen Untersuchungsdesign ist weder das technische noch das soziale Subsystem exklusiv beteiligt. Vielmehr sind es Wechselbeziehungen zwischen den sozialen und technischen Subsystemen beziehungsweise Wechselbeziehungen innerhalb dieser, die Daten erstellen. Die Datenerstellung innerhalb des Gesamtsystems ist ein nicht-linearer Prozess, dessen Vorgaben und konkrete Ausgestaltung nicht nur die entstehenden Daten, sondern auch deren Qualität bestimmt.

Die Entstehung von fehlenden Daten, auch durch Löschung fehlerhafter Daten, wird maßgeblich durch zwei Komponenten bestimmt. Zum einen ist das vorgegebene Untersuchungsdesign, welches im nachfolgenden Abschnitt 2.1.1 ausführlich erläutert wird, wichtig. Zum anderen ist die vorherrschende Kommunikationsstruktur, welche in Abschnitt 2.1.2 diskutiert wird, innerhalb des soziotechnischen Systems relevant. Letztendlich bestimmen die involvierten Akteure, ob Mensch oder Maschine, deren Interaktion und deren Befassung mit sich selbst[1], ob fehlende oder fehlerhafte Werte entstehen.

2.1.1 Untersuchungsdesign

Die wohl wichtigste Frage bei der Ausfallursache ist, ob das Untersuchungsdesign fehlende Werte vorsieht. Das gewählte Untersuchungsdesign kann unmittelbare Wirkung auf alle Stufen des datenanalytischen Untersuchungsprozesses entfalten und wird zu Beginn, wenn auch zum Teil implizit, festgelegt. Im Hinblick auf das Auftreten von fehlenden Werten kann zwischen jenen Untersuchungsdesigns unterschieden werden, die fehlende Werte per Design vorsehen und jenen, bei denen fehlende Werte nicht vorgesehen sind. Sind fehlende Werte per Design vorgesehen, kann unterschieden werden zwischen impliziter und expliziter Planung fehlender Werte. Sind fehlende Werte nicht im Untersuchungsdesign vorgesehen, kann zwischen einem fehler- und einem mangelhaften Untersuchungsdesign unterschieden werden (vgl. Bankhofer, 1995, S. 8). Beispiele für alle vier Kombinationen sind im Folgenden gegeben.

[1] Akzeptiert ein Mensch beispielsweise die ihm zugeordnete Rolle im datenanalytischen Untersuchungsprozess oder bestimmte Komponenten dieses Prozesses nicht (vgl. Müllerleile und Nissen, 2014, S. 179–180), reduziert die Befassung der Person mit sich selbst zwangsweise die Datenqualität.

Beispiel 2.1: *Explizit geplante fehlende Werte*
Fehlende Werte können explizit von einem Untersuchungsdesign vorgesehen werden. Eine Möglichkeit zur gezielten Nutzung fehlender Daten, um Informationen zu nominal einem Drittel mehr Merkmalen zu erheben, wird von Graham et al. (2006) diskutiert. Die Autoren schlagen vor, Fragebögen in einem „three-form design“ zu konzipieren. Wie der Name andeutet, werden aus einem Grundtyp des Fragebogens drei Fragebögen abgeleitet. Bei den abgeleiteten Fragebögen fehlt jeweils ein anderes Drittel der Fragen. Da die Fragebögen zufällig und gleichmäßig auf die Befragten verteilt werden, entsteht hierdurch ein Ausfallmechanismus, mit dem relativ einfach umzugehen ist.

Beispiel 2.2: *Implizit geplante fehlende Werte*
Ein vom Untersuchungsdesign vorgesehener Ausfall kann sich auch implizit ergeben, je nach Datenerhebungsmethode. So kann es bei der Zusammenführung mehrerer Sekundärquellen immer passieren, dass gewisse Merkmalskombinationen niemals gemeinsam beobachtet werden. Dieser Zustand kann auch bei Primärdaten auftauchen; beispielsweise bei der Verwendung gewisser Filterfragen beziehungsweise Verzweigungslogiken in einem Fragebogen. Eine Filterfrage liegt vor, wenn eine folgende Frage nur gestellt wird, sofern voher eine bestimmte Antwort gegeben wurde. Soll die Frage nach der Lieblingsgummibärfarbe nur beantwortet werden, wenn vorher festgestellt wurde, dass der Befragte Gummibären grundsätzlich konsumiert, handelt es sich bei der zuerst gestellten Frage um eine Filterfrage.

Beispiel 2.3: *Ungeplante fehlende Werte wegen fehlerhaften Designs*
Fehlende Werte können auch ungeplant wegen eines fehlerhaften Designs auftreten. So wird in diesem Fall das Untersuchungsdesign in einer Art und Weise festgelegt, dass es zwangsweise bei der anschließenden Datenerhebung zu fehlenden Werten kommt (vgl. Bankhofer, 1995, S. 8). Als Beispiel kann hier das Merkmal „Lieblingsbier“ verwendet werden. Wird dieses Merkmal bei Personen, die kein Bier konsumieren, erhoben, treten zwangsweise fehlende Werte auf. Praktisch ist zwischen diesem Fall und dem der implizit geplanten fehlenden Werte schwer zu unterscheiden. Dabei ist die Bedeutung des Wortes „fehlerhaft“ und die Existenz besserer Designvarianten zu berücksichtigen. Für den

korrekten Umgang mit den fehlenden Werten hat diese Differenzierung keine Auswirkung.

Beispiel 2.4: *Ungeplante fehlende Werte wegen mangelhaften Designs*
Wird das Untersuchungsdesign in einer Art festgelegt, dass es bei der Datenerhebung unter Umständen zu fehlenden Werten kommt, entstehen diese ungeplanten fehlenden Werte wegen eines mangelhaften Designs (vgl. Bankhofer, 1995, S. 8). Ein mangelhaftes Untersuchungsdesign kann vorliegen, wenn beispielsweise Fragen in einem Fragebogen kompliziert oder nicht eindeutig formuliert sind. Ferner ist ein Design mangelhaft, wenn zur Beantwortung der Fragen nicht hinreichend Zeit alloziert wird, beziehungsweise wenn zu viele Fragen für die geplante Zeit vorgesehen werden. Hier werden die am Ende stehenden Fragen häufiger unbeantwortet bleiben, wodurch sich ein monotones Ausfallmuster ergeben kann.

2.1.2 Kommunikationsstruktur

Die Ursache des Datenausfalls kann entweder im betrachtenden oder in dem betrachteten System inhärent sein. Da Daten immer über ein System von einem anderen System erhoben werden müssen, kann der Datenausfall durch diese Systeme oder die Interaktion dieser erzeugt werden. Zur Erklärung und Einordnung, wo in der Datenerhebung und -aufbereitung fehlende Werte entstehen können, kann das Shannon-Weaver-Modell (Shannon und Weaver, 1949) verwendet werden. Abbildung 2.3 zeigt eine abgewandelte, übersetzte Darstellung des schematischen Diagramms von Shannon und Weaver (1949, S. 34).

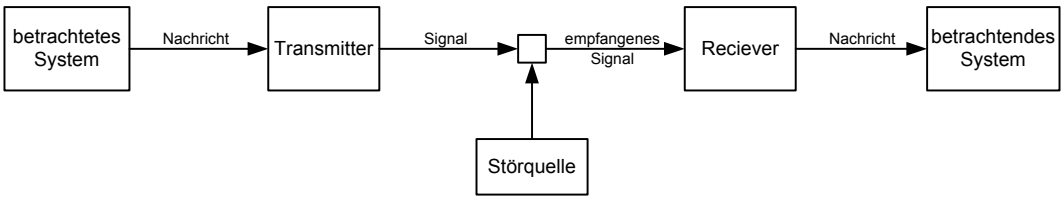


Abbildung 2.3: Das Shannon-Weaver-Modell (in Anlehnung an Shannon und Weaver, 1949, S. 34)

Das Modell besteht grundsätzlich aus fünf Teilen, wobei in jedem dieser Teile Gründe für fehlende Werte beziehungsweise fehlerhafte Werte, deren Korrektur

in der Datenaufbereitungsphase zu fehlenden Werten führt, vorhanden sein können:

1. Das *betrachtete System* erstellt eine Serie an Nachrichten, enthält Informationen oder weist sonstige Eigenschaften auf. Im datenanalytischen Prozess muss das betrachtete System nicht zwangsweise auch das Untersuchungsobjekt sein. Handelt es sich um die Datenaufbereitungsphase, kann das betrachtete Systen aus jenem Datenmaterial, mittels dem eine auswertbare Datenmatrix erstellt werden soll, bestehen.

   Die folgenden Ausfallursachen sind an dieser Stelle von besonderer Bedeutung:

   - **Obsoleszenz:** Das betrachtete System stellt Daten zur Verfügung, die aufgrund ihres Alters als (teilweise) unbrauchbar einzuschätzen sind. Dies kann sowohl bei Sekundär- als auch bei Primärdaten der Fall sein, wenn der Zeitraum zwischen Erhebung und Auswertung zu groß wird.
   - **Integration:** Wenn Daten aus mehreren Quellen integriert werden, kann es vorkommen, dass die verschiedenen Quellen nicht zu den gleichen Objekten Merkmale zur Verfügung stellen. Dies tritt insbesondere dann auf, wenn eine oder mehrere Primärdatenquellen mit verschiedenen Sekundärdatenquellen zusammengeführt werden.
   - **Nichtexistenz:** Bestimmte Informationen über das betrachtete System existieren nicht. Dies kann der Fall sein, wenn bestimmte Werte nicht eingenommen werden können beziehungsweise nicht erhoben wurden. Ist die Datenquelle ein Mensch, so kann es sein, dass er trotz redlichem Bemühen zu bestimmten Fragen keine Antwort weiß. Die Fragen beziehen sich beispielsweise auf längst vergangene oder als unwichtig bewertete Ereignisse. Ist die Datenquelle eine Datenbank, kann es sich hierbei um nicht gepflegte Felder handeln oder um jene Einträge, die aus Datenschutzgründen gelöscht wurden.
   - **Unzugänglichkeit:** Die Daten über das betrachtete System können unzugänglich sein. Bei einer Befragung kann dies auf Antwort-

verweigerung, bei einer Datenbankabfrage kann dies auf mangelnde Rechte zurückzuführen sein.

2. Der *Transmitter* codiert die Nachrichten in einer Art und Weise, so dass diese über das entsprechende Medium übertragen werden können. Bei einer verbalen Befragung werden beispielsweise die Gedanken der befragten Person als Sprache ausgegeben. Bei einem Telegraphen werden Nachrichten in eine Sequenz von kurzen und langen elektrischen Impulsen codiert. Nachrichten und Signale im Modell müssen nicht zwangsweise aktiv sein. Bei einer reinen Beobachtung kann ein Unterlassen durchaus ein Signal darstellen. Es bietet sich an, folgende Ausfallursachen an dieser Stelle zu erwähnen:

   - **Nicht-Codierung:** Der Transmitter codiert die an ihn gerichtete Nachricht nicht. Dies wäre der Fall, wenn ein Temperatursensor an einem Hochofen ausfällt und keinen Wert angibt. Auch liegt ein Fall der Nicht-Codierung vor, wenn ein Proband eine Frage liest und beantworten will, zwischenzeitlich jedoch abgelenkt wird und infolge der Ablenkung die Frage überspringt.
   - **Fehlerhafte Codierung:** Sollte der Transmitter Werte fehlerhaft codieren, welches nachher erkannt wird, können sich fehlende Werte durch Löschung der fehlerhaften Werte ergeben. Eine fehlerhafte Codierung kann vorliegen, wenn aufgrund defekter Sensorik die an einer Photovoltaik-Anlage gemessene Spannung immer den gleichen Wert angibt, obwohl tatsächlich Spannungsschwankungen zu verzeichnen sind. Fehlerhafte Codierung liegt zudem vor, wenn ein ausländischer Student in seiner Muttersprache auf die ihm gestellten Fragen antwortet. Kann der Receiver diese nicht verstehen, muss diese Sprachbarriere als fehlerhafte Codierung interpretiert werden.

3. Der genutzte *Kanal* ist lediglich das Medium, welches zur Signalübertragung verwendet wird. Dies können Kabel, Luft oder Papier sein. Während der Übertragung kann das Signal gestört werden, so dass das übertragene Signal nicht dem empfangenen entspricht. Eine Störung an dieser Stelle kann auf zwei Arten fehlende Werte erzeugen:

- **Unterbrechung:** Das vom Transmitter gesendete Signal kann auf dem Weg zum Receiver unterbrochen werden, wobei es zu einem Totalausfall des Signals kommt. Dies kann zum einen der Fall sein, wenn bei der Überwachung mit einem technischen System die Stromversorgung zu diesem unterbrochen wird. Zum anderen kann dies auch der Fall sein, wenn die Sekretärin Kaffee über ausgefüllte Fragebögen verschüttet. Zerstört der verschüttete Kaffee zumindest teilweise die Lesbarkeit der Antworten, wurde das Signal, in Form des Fragebogens, unterbrochen.
- **Beeinflussung:** Bei der Übertragung des Signals über einen bestimmten Kanal können fehlerhafte Werte entstehen, indem das Signal beeinflusst wird. Werden diese fehlerhaften Werte wiederum in Folge erkannt, kann es zu fehlenden Werten kommen. Eine Beeinflussung des Signals liegt vor, wenn ein Tintenfass voll königsblauer Tinte umgestoßen wird und zufälligerweise aus einer Drei eine Acht macht. Des Weiteren kann ein Signal beeinflusst werden, wenn das Medium der Übertragung den Menschen beinhaltet. Werden die Antworten eines Fragebogens in ein maschinenlesbares Format von einem Menschen übertragen, können sich durch Unachtsamkeit oder mangelhafte Fertigkeiten Fehler ergeben. Zudem kann eine Beeinflussung des Signals vorliegen, wenn sich mehrere Signale ein Übertragungsmedium teilen. Wird das Medium nicht korrekt auf die Signale aufgeteilt, z. B. Aufteilung mehrerer frequenzmodulierter Signale auf überlappende Frequenzbänder beim Frequenz-Multiplexing, werden die Signale zerstört und können vom Receiver nicht mehr korrekt decodiert werden.

4. Der *Receiver* führt in der Regel die Gegenoperation zum Transmitter durch. Er rekonstruiert aus dem empfangenen Signal die Nachricht. Ähnlich wie beim Transmitter sind die folgenden zwei Ausfallgründe erwähnenswert:

   - **Nicht-Decodierung:** Basierend auf dem empfangenen Signal und der Struktur des Receivers ist es möglich, dass der Reciever das an ihn weitergeleitete Signal nicht decodieren kann oder dieses nicht

decodiert wird. Wird eine Beobachtung mittels eines Fernglases vorgenommen, kann ein Beschlagen der Linsen mit Kondenswasser dazu führen, dass wichtige Sachverhalte nicht gesehen werden. Im gleichen Zuge kann das Tag-Träumen einer wissenschaftlichen Hilfskraft dazu führen, dass sie einen Sachverhalt hätte aufnehmen können. Die Ablenkung führte in diesem Fall jedoch dazu, dass das optisch Wahrgenommene nicht decodiert wird. Sollte der Reciever Teil eines technischen Systems sein, ist es denkbar, dass ein durch eine Störung beeinflusstes Signal hinreichend verzerrt ist, so dass eine Decodierung nicht möglich ist.

- **Fehlerhafte Decodierung:** Sollte der Receiver Werte fehlerhaft decodieren, welches nachher erkannt wird, können sich fehlende Werte durch Löschung der fehlerhaften Werte ergeben. Eine fehlerhafte Decodierung kann beispielsweise dann stattfinden, wenn Unklarheit über die verwendete Gestik herrscht. Ein Wissenschaftler aus Deutschland, der sich vorher nicht mit bulgarischer Kultur auseinandergesetzt hat, wird das Nicken eines Befragten als Zustimmung fehlinterpretieren. Eine Quelle von fehlenden Werten durch fehlerhafte Decodierung kann sich durch die Fehlinterpretation von Steuerzeichen in einer Computerdatei ergeben. Ein bekanntes Dateiformat, die CSV (comma separated values), weist Unterschiede je nach Kulturraum auf. Ist, wie im angloamerikanischen Raum üblich, ein Punkt das Dezimalzeichen, werden unterschiedliche Merkmalsausprägungen bei diesem Dateiformat mit einem Komma getrennt. Ist ein Komma das Dezimalzeichen, wird mit einem Semikolon getrennt. Durch eine fehlerhafte Decodierung kann sich unmittelbar für jedes Objekt eine unterschiedliche Anzahl an Merkmalen ergeben.

5. Das *betrachtende System* stellt die Person oder Sache dar, für welche die Nachrichten, Informationen oder sonstigen Eigenschaften des betrachteten Systems von Interesse sind. Im datenanalytischen Prozess muss das betrachtende System nicht zwangsweise auch das Untersuchungssubjekt sein. Wird die Datenerhebungsphase betrachtet, kann es sich bei dem betrachtenden System um den Befragten, dem gewisse Fragen gestellt

werden, handeln. Die folgenden Ausfallursachen sind an dieser Stelle von besonderer Bedeutung:

- **Nicht-Aufzeichnung:** Die Nachricht wurde zwar erfolgreich vom Receiver decodiert, doch das betrachtende System entscheidet sich, die Daten nicht aufzuzeichnen. Gründe für eine solche Nicht-Aufzeichnung können mannigfaltig sein, sind jedoch meist auf Fehlfunktionen zurückzuführen. Zwar ist insbesondere das Versagen von technischen Systemen an dieser Stelle plausibel, aber auch menschliches Versagen ist denkbar. So kann es der Fall bei einer Kamera sein, dass zwar das optische Signal erfolgreich decodiert wird, jedoch der Speicher in bestimmten Sektoren fehlerhaft und nicht-beschreibbar ist. Hierdurch kann das empfangene Signal nicht oder nur teilweise gespeichert werden. Das Pendant, wenn das betrachtende System ein Mensch ist, wäre, wenn ein Beobachter abgelenkt beziehungsweise unaufmerksam ist. So sitzt der Beobachter zwar vor dem Untersuchungsobjekt, nimmt jedoch wegen Tag-Träumen das Gesehene nicht wahr.
- **Abweisung:** Ein vom Receiver erfolgreich decodiertes Signal kann von dem betrachtenden System abgewiesen werden. Nachrichten werden zumeist abgewiesen, wenn der Inhalt fehlerhaft ist. So besteht die Möglichkeit, dass zwar das empfangene Signal richtig decodiert wird, der Inhalt der Nachricht aber nicht den Anforderungen des betrachtenden Systems entspricht. Anforderungen des Systems können unterschiedlichster Art sein. Sie reichen von objektiv nachvollziehbaren Kriterien bei Edit-Systemen, die festlegen, dass bestimmte Merkmalskombinationen nicht auftreten dürfen, bis hin zu Motivationsproblemen bei einem Beobachter. Dies kann sein, wenn das Signal aufgrund von Störungen nicht dem ursprünglichen Signal entspricht. Erkennt das betrachtende System, dass die Nachricht unsinnig ist, so kann die Entscheidung zur Abweisung getroffen werden.

Praktisch gesehen ist es nicht nur schwierig, zwischen betrachtetem System und Transmitter beziehungsweise zwischen Receiver und betrachtendem

System zu trennen, sondern auch zwischen betrachtetem und betrachtendem System. Zum einen liegt dies daran, dass es sich in der Realität häufig um Aneinanderreihungen von Kommunikationssystemen handelt, bei denen Anfangs- und Endpunkt schlecht definiert sind. Zum anderen liegt dies an der verwendeten Aggregationsstufe der Betrachtung. Beispielsweise mag das Internet auf einer hohen Aggregationsstufe als Übertragungsmedium interpretiert werden, auf niedrigeren Aggregationsstufen ergibt sich eine lange und mannigfaltige Verkettung an Transmittern und Receivern. Auch ist es bei einer Befragung schwer zu beurteilen, ob der Befragte inkonsistente Antworten gegeben hat, weil er die Fragen nicht verstanden hat, oder weil es dem Befragten an Ausdrucksvermögen mangelt.

2.2 Muster fehlender Daten

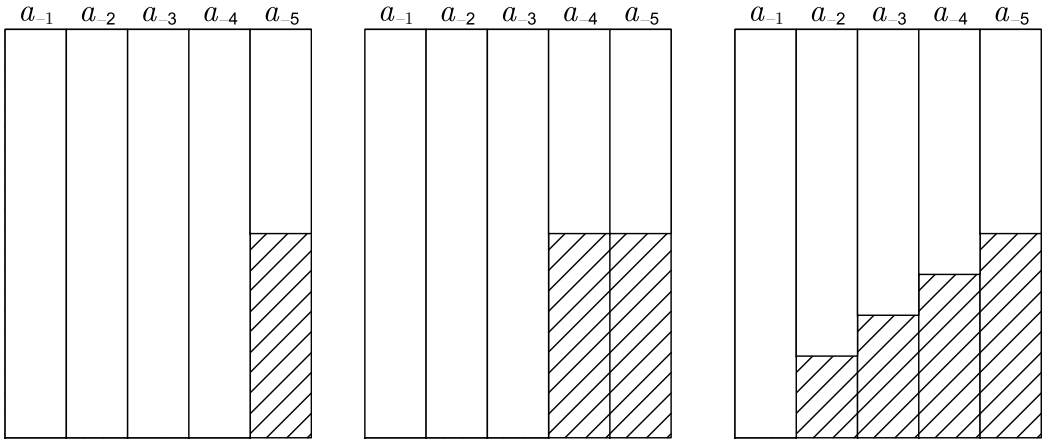
Es ist sinnvoll, zwischen dem Muster der fehlenden Daten und dem Ausfallmechanismus zu unterscheiden. Während Ausfallmechanismen die bestehenden Beziehungen zwischen dem Vorhandensein von Daten und den Ausprägungen von Variablen charakterisieren, beschreiben die Ausfallmuster lediglich, wo fehlende Werte in der Datenmatrix auftreten (vgl. Little und Rubin, 2002, S. 4 beziehungsweise Enders, 2010, S. 2).

Gegeben sei die Datenmatrix $A = (a_{ik})_{n,m}$. Die Merkmalsausprägungen des i-ten Objekts entsprechen dem Vektor $a_{i-} = (a_{i1}; \ldots; a_{im})^T$ und die Merkmalsausprägungen aller Objekte des k-ten Merkmals werden im Vektor $a_{-k} = (a_{1k}; \ldots; a_{nk})^T$ zusammengefasst. Ferner sei $V = (v_{ik})_{n,m}$ die Indikatormatrix der fehlenden Ausprägungen (missing-data indicator matrix, MD-Indikatormatrix) mit $v_{ik} = 1$, wenn a_{ik} fehlt und $v_{ik} = 0$ wenn a_{ik} vorhanden ist. Dann definiert die Struktur von V das Muster der fehlenden Daten[2]. Sechs Strukturen sind prototypisch in den Abbildungen 2.4 und 2.5 dargestellt.

Das in Abbildung 2.4a dargestellte univariate Ausfallmuster ist das simpelste Muster, das bei Vorhandensein von fehlenden Werten existieren kann. Hier beschränken sich die fehlenden Werte auf eine einzelne Variable. Dieses Muster tritt insbesondere bei geplanten Experimenten auf, bei denen lediglich eine

[2] Zur Erkennung von Mustern ist gegebenenfalls eine geschickte objekt- oder merkmalsweise Sortierung von V erforderlich (vgl. Little und Rubin, 2002, S. 4).

Variable (in diesem Fall a_{-5}) als Ergebnis des Experiments analysiert werden soll und die restlichen, vollständigen Variablen direkt durch den Forscher manipulierbar sind. Auch ist das univariate Muster interessant, weil jedes der anderen Muster auf eine Serie von Konstellationen dieses Musters zurückgeführt werden kann. Eine Beispielsituation des geplanten Experiments, bei dem ein univariates Ausfallmuster auftreten kann, kommt aus der Landwirtschaft (vgl. Little und Rubin, 2002, S. 4). Hier sind häufig die Abhängigkeiten zwischen dem Ertrag eines Feldes (z. B. Tonnen Mais) und mehreren nominal skalierten Variablen (wie etwa dem verwendeten Dünger oder der Art der Aussaat) von Interesse. Da die unabhängigen Faktoren vollständig beobachtet oder gar im Voraus festgelegt werden, können lediglich in der abhängigen Variable fehlende Werte entstehen (beispielsweise, weil die Keimung des Saatguts ausbleibt).



(a) Univariat (b) Unit-Nonresponse (c) Monoton

Abbildung 2.4: Beispiele für Muster fehlender Daten; Schattierung bedeutet $v_{ik} = 1$ (in Anlehnung an Little und Rubin, 2002, S. 5)

Ein spezielles multivariates Muster ist in Abbildung 2.4b dargestellt. Hierbei handelt es sich um das bei Umfragen häufig auftretende Muster Unit-Nonresponse. Unit-Nonresponse liegt genau dann vor, wenn sämtliche Merkmale bei bestimmten Objekten fehlen, die über die a priori vorhandenen Informationen beziehungsweise vorher festgelegten Designparameter hinausgehen. Das Unit-Nonresponse-Muster tritt genau dann auf, wenn eine Untermenge der Befragten den Fragebogen nicht oder nicht vollständig ausfüllt, wobei das Nicht-Ausfüllen des Fragebogens auf Antwortverweigerung oder einen der anderen in Abschnitt 2.1 aufgeführten Gründe zurückzuführen ist. Dieses Muster

kann zudem in geplanter Form beim Double-Sampling (vgl. Hamaker und van Strik, 1955) auftreten, wenn einige, günstig zu erhebende Merkmale bei der gesamten Stichprobe erhoben werden, und wenige, teuer zu erhebende Merkmale bei einer kleineren Untermenge der Stichprobe gemessen werden. Das Gegenstück zu Unit-Nonresponse ist das Item-Nonresponse. Bei Item-Nonresponse liegt kein grundsätzlicher Datenausfall vor, sondern nur ein Fehlen einzelner Merkmalsausprägungen.

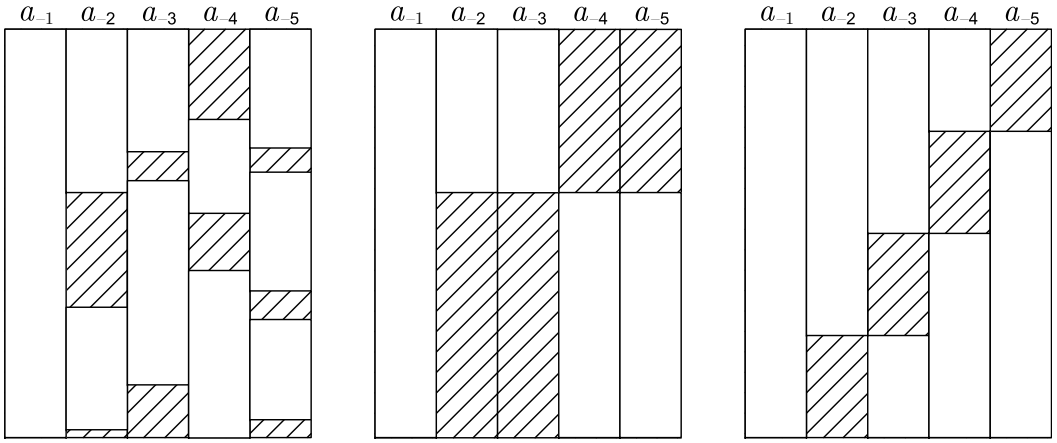
In Abbildung 2.4c ist ein monotones Ausfallmuster angedeutet. Monotone Ausfallmuster liegen vor, wenn eine Permutation der verschiedenen Merkmalsvektoren $a_{-1}, \ldots, a_{-m}$ existiert, so dass, wenn bei a_{-l} ein Wert fehlt, dieser auch bei a_{-k} für alle $k = l + 1, \ldots, m$ fehlt (vgl. Tang et al., 2003, S. 759). Dieses Ausfallmuster tritt insbesondere bei Longitudinalstudien auf, wenn Probanden oder andere Untersuchungsobjekte vor Ende der Studie ausscheiden ohne wieder einzutreten[3]. Visuell ähnelt dieses Muster, bei einer Sortierung der Variablen, Stufen, da der Anteil der fehlenden Werte, der in jedem Merkmal vorhanden ist, monoton ansteigt. Das Vorliegen dieses Musters hat den Vorteil, dass sich die Komplexität einiger beliebter Methoden zum Umgang mit fehlenden Werten deutlich reduziert. Ein monotones Muster reduziert zum Beispiel die Rechenzeit des EM-Algorithmus (Dempster et al., 1977), da keine Iterationen über die Merkmale benötigt werden (Schafer, 1997, S. 218–238).

Das in der Praxis wohl am häufigsten anzutreffende Ausfallmuster ist das allgemeine Muster fehlender Werte. Wie Abbildung 2.5a andeutet, treten die fehlenden Werte wahllos über die Datenmatrix verteilt auf. Zwar genügt die Beschreibung eines allgemeinen Musters einem zufälligen Auftreten von fehlenden Werten, jedoch nur bezüglich ihrer Position in der Datenmatrix. Durchaus kann der Ausfallmechanismus weiterhin systematischer Natur sein. Allgemeine Muster können immer auf eine Reihe der anderen diskutierten Muster zurückgeführt werden. Im Extremfall kann, sofern die Objekt- und Merkmalsreihenfolge beliebig ist[4], ein allgemeines Muster auf maximal m univariate Muster zurückgeführt werden.

---

[3] Dieser Umstand ist auch als Attrition bekannt.

[4] Dies ist immer der Fall, wenn die Informationen, die in der Objektreihenfolge vorhanden sind, in ein weiteres Merkmal codiert werden können, und die Merkmalsreihenfolge keine Informationen enthält.



(a) Allgemein (b) Dateizuordnung (c) Teilbefragung

Abbildung 2.5: Beispiele für Muster fehlender Daten; Schattierung entspricht $v_{ik} = 1$ (in Anlehnung an Enders, 2010, S. 4)

Ein insbesondere in der Informatik auftretendes Ausfallmuster ist das Muster, das sich bei dem Zusammenfügen mehrerer Dateien ergeben kann (vgl. Abbildung 2.5b). Dieses Dateizuordnungsmuster (file matching, vgl. Little und Rubin, 2002, S. 5) entsteht, wenn bestimmte Merkmale niemals für bestimmte Gruppen an Objekten gemeinsam erhoben werden. Dieses Muster kann bei Abfragen von relationalen Datenbanken auftreten, wenn beispielsweise Tabellen einer SQL-Datenbank mit einer *outer join* (vgl. Codd, 1979, S. 406 f.)[5] Operation zusammengefügt werden. Auch tritt dieses Muster häufig auf, wenn Daten, insbesondere Sekundärdaten, aus heterogenen Quellen für eine Analyse zusammengeführt werden.

Ein weiteres Ausfallmuster (vgl. Abbildung 2.5c), bei dem der Ausfall ähnlich wie beim Double-Sampling per Design festgelegt wird, ist die Teilbefragung (partial questionnaire design, Wacholder et al., 1994, S. 623 ff.; beziehungsweise split questionnaire, Raghunathan und Grizzle, 1995, S. 54 ff.). Bei dieser speziellen Form des Matrix-Samplings (vgl. Raghunathan und Grizzle, 1995, S. 54) wird nicht jedem Befragten jede Frage gestellt, um die sogenannte Befragungslast (response burden) zu reduzieren. Ein Teil der Fragen wird zwar weiterhin allen Personen vorgelegt, die verbleibenden Fragen werden jedoch nie alle von derselben Gruppe an Befragten beantwortet. Entsprechend der

[5] Unterschiedliche Muster ergeben sich bei Verwendung von *full outer join* oder *left* beziehungsweise *right outer join*.

verwendeten Zuordnungskriterien können hier per Design Ausfallmechanismen erzeugt werden, deren Vorliegen erstrebenswert für den späteren Umgang mit den fehlenden Werten ist. Je nach Fragen und Zielstellung kann die Belastung der Befragten um bis zu einem Anteil von Eins durch die Anzahl der Gruppen reduziert werden. Alternativ lässt sich hierdurch die Anzahl an Fragen um die entsprechende Proportion erhöhen (vgl. Graham, 2009, S. 566).

Historisch waren Ausfallmuster insbesondere vor der Entwicklung des Konzeptes des Ausfallmechanismus relevant (Enders, 2010, S. 5). Ältere, kaum noch in der Literatur erwähnte Methoden zum Umgang mit fehlenden Werten, setzten für die Anwendbarkeit das Vorliegen bestimmter Ausfallmuster voraus (vgl. Little und Rubin, 2002, S. 133 ff.). Praktisch relevant kann die Analyse der Struktur von V dennoch sein, da mittels der vorhandenen Ausfallmuster potenzielle Gründe des Ausfalls identifiziert werden können.

2.3 Ausfallmechanismen

Von zentraler Bedeutung für die Wahl einer Missing-Data-Methode ist die korrekte Identifizierung des vorhandenen Ausfallmechanismus. Diese Mechanismen stellen Beschreibungen des Zufallsprozesses dar, der die Verteilung der MD-Indikatormatrix bestimmt. Mechanismen beschreiben somit die Gründe, im Sinne eines zugrundeliegenden stochastischen Prozesses, warum Werte fehlen. Insbesondere definieren sie die Beziehungen innerhalb der Daten, und nicht die Muster, die sich letztendlich ergeben. Rubin (1976a) beschrieb erstmals die bedingte Verteilung von V mit Parameter ϕ als Bernoulli-Prozess und stellte die notwendigen Bedingungen auf, dass ein Mechanismus, der dem Erscheinen von fehlenden Werten zugrunde liegt, vernachlässigbar ist. Vernachlässigbar bedeutet in diesem Kontext jedoch nicht, dass die explizite Berücksichtigung der fehlenden Daten vernachlässigbar ist, sondern nur, dass eine Modellierung des Ausfallmechanismus entbehrlich bleibt.

Grundlage der Diskussion von Ausfallmechanismen bildet die in Abbildung 2.6 dargestellte Unterteilung der Datenmatrix A in beobachtete und unbeobachtete Teile, A^{obs} und A^{mis}, sowie deren Beziehungen[6]. Geht man davon aus,

[6] A^{obs} und A^{mis} bezeichnen vorhandene und fehlende Realisierungen der Merkmalsausprägungen. Sie stellen keine Matrizen dar (vgl. Bankhofer, 1995, S. 6).

dass die Daten Zeitpunkt gebunden sind, d.h. nicht nacherhoben werden können, so handelt es sich bei A^{obs} um die beobachtbaren und bei A^{mis} um die nicht beobachtbaren Teile von A, aus denen V abgeleitet wird.

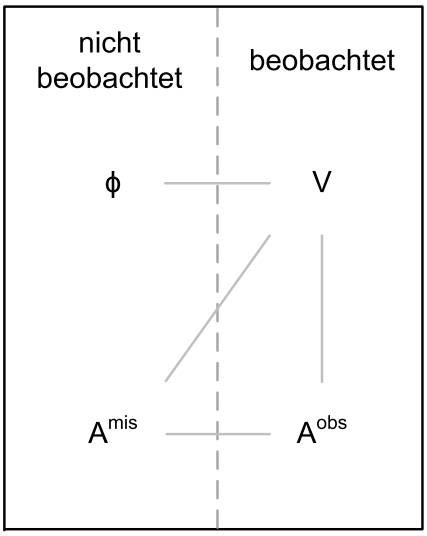


Abbildung 2.6: Mögliche Beziehungen im Missing-Data-Modell (in Anlehnung an Schafer und Graham, 2002, S. 152)

Der Aufbau des folgenden Abschnittes richtet sich im Wesentlichen nach der heute gängigen Aufteilung von Ausfallmechanismen, wie sie auch in Little und Rubin (2002, S. 11–12) zu finden ist. Zunächst werden einzeln jene Ausfallmechanismen, die zur Wahl eines geeigneten Verfahrens zur Behandlung fehlender Daten relevant sind, beschrieben. Danach erfolgt in Abschnitt 2.3.4 eine Diskussion über die Plausibilität des Not Missing at Random Mechanismus in Untersuchungen der Sozialforschung.

2.3.1 Missing Completely at Random (MCAR)

Die fehlenden Werte einer unvollständigen Datenmatrix werden als missing completely at random oder kurz MCAR bezeichnet, wenn für die Verteilung f der MD-Indikatormatrix V folgende Bedingung gilt:

$$f\left(V|A;\phi\right) = f\left(V|\phi\right) \ \forall\, A, \phi. \tag{2.5}$$

Dies bedeutet, dass die Verteilung des Auftretens fehlender Werte unabhängig von den konkreten Werten der Datenmatrix ist, egal, ob diese beobachtet oder nicht beobachtet wurden. Abbildung 2.7 verdeutlicht, welche Abhängigkeiten

vorliegen dürfen, so dass die Bezeichnung MCAR für den Ausfallmechanismus zutrifft. Beziehungen dürfen lediglich zwischen ϕ und V sowie A^{obs} und A^{mis} bestehen. Anzumerken ist, dass an dieser Stelle keine Aussagen über die vorliegenden Ausfallmuster gemacht werden.

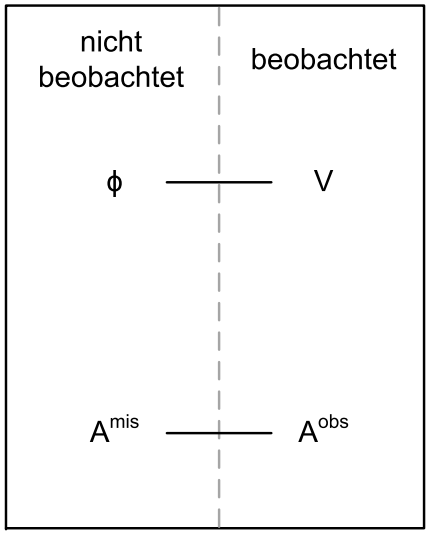


Abbildung 2.7: MCAR-Beziehungen im Missing-Data-Modell (in Anlehnung an Schafer und Graham, 2002, S. 152)

Beispiel 2.5: *MCAR-Ausfall bei einem Merkmal*
Gegeben sei ein Satz vollständiger Merkmalsvektoren A^{cv} und ein unvollständiger Merkmalsvektor a_{-1}^{mv}. So liegt ein MCAR-Ausfallmechanismus vor, wenn:

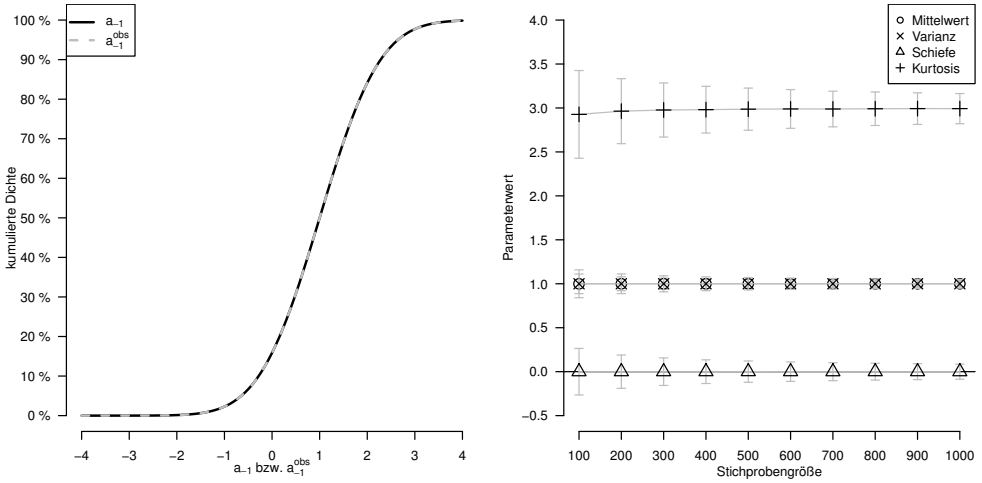
$$Pr\left(v_{i1}^{mv} = 1 | a_{i-}; \phi\right) = \text{konst } \forall i = 1, \ldots, n. \tag{2.6}$$

Liegt ein MCAR-Mechanismus vor, so handelt es sich bei den beobachteten Werten A^{obs} um eine einfache Stichprobe der gewünschten Daten (vgl. Enders, 2010, S. 7), da der Selektionsmechanismus in keiner Beziehung zu den Daten steht. Somit ist auch klar, dass im Fall von MCAR-Daten eine Analyse von A^{obs} valide Erkenntnisse über A liefert, und lediglich die Konsequenzen eines reduzierten Stichprobenumfangs hinzunehmen sind. Der MCAR-Ausfallmechanismus ist insbesondere plausibel, wenn das Fehlen der Daten geplant ist (vgl. Beispiel 2.7). MCAR ist jedoch eine unrealistisch starke Annahme, wenn das Fehlen der Daten nicht geplant ist, da das Fehlen dieser Werte meist einen Grund hat, der in den beobachteten Werten zu finden ist (vgl. Dillman et al., 2002, S. 18).

Beispiel 2.6: *Auswirkungen von MCAR-Ausfall auf eine Normalverteilung*
Gegeben sei ein normal verteilter Merkmalsvektor a_{-1} mit Erwartungswert $\mu = (1)$ und Varianz $\sigma_{11} = 1$, betroffen von dem folgenden MCAR-Ausfallmechanismus:

$$Pr\left(v_{i1} = 1\right) = 0{,}194 \;\; \forall i = 1, \ldots, n. \tag{2.7}$$

So beträgt $\tilde{v}^{mis} = \tilde{v}^{mis}_{\bullet 1}$, unabhängig von jeglichen anderen Parametern und unbeobachteten Variablen, circa 19,4%[7]. Deutlich zu erkennen ist, dass sich die Verteilungen von a^{obs}_{-1} und a_{-1} nicht unterscheiden (vgl. Abbildung 2.8a). Mittelwert, Varianz, Schiefe und Exzess von a_{-1} können erwartungstreu aus a^{obs}_{-1} geschätzt werden[8] (vgl. Abbildung 2.8b).



(a) Verteilungsfunktionen (b) Parameterschätzungen

Abbildung 2.8: Vergleich von a_{-1} und a^{obs}_{-1} bei MCAR-fehlenden Daten

Beispiel 2.7: *MCAR-Ausfallmechanismus durch Double-Sampling*
Eine Situation, bei der das Fehlen von Daten als MCAR geplant ist (vgl. Beispiel 2.1), ist das sogenannte Double-Sampling (vgl. Statistical Research Group, 1948; Hamaker und van Strik, 1955). Beim Double-Sampling werden zunächst eine Stichprobe von der Grundgesamtheit gezogen und etliche einfache Variablen erhoben. Danach werden für eine kleinere zufällige Teilmenge der Stichprobe weitere Merkmale aufgezeichnet, deren Erhebung vergleichswei-

[7] Wobei der Wert 19,4% willkürlich zu Demonstrationszwecken gewählt wurde.

[8] Ergebnisse der Auswertung von 100.000 simulierten Datensätzen pro Stichprobengröße.

se teuer ist. Das resultierende Muster entspricht dem Unit-Nonresponse (vgl. Abbildung 2.4b).

2.3.2 Missing at Random (MAR)

Die Daten, ausgehend von einer unvollständigen Datenmatrix, werden als missing at random (kurz MAR) bezeichnet, wenn folgendes für die Verteilung von V gilt:

$$f\left(V|A;\phi\right) = f\left(V|A^{obs};\phi\right) \;\forall\, A^{mis},\phi. \tag{2.8}$$

In Worten bedeutet dies, dass das Auftauchen von fehlenden Werten rein durch beobachtete Werte und den Störparameter ϕ (nuisance parameter, vgl. Rüger, 2002, S. 7) bedingt ist. Die Ausfallwahrscheinlichkeit ist also unabhängig von jeglichen fehlenden Werten, wie es in Abbildung 2.9 verdeutlicht wird.

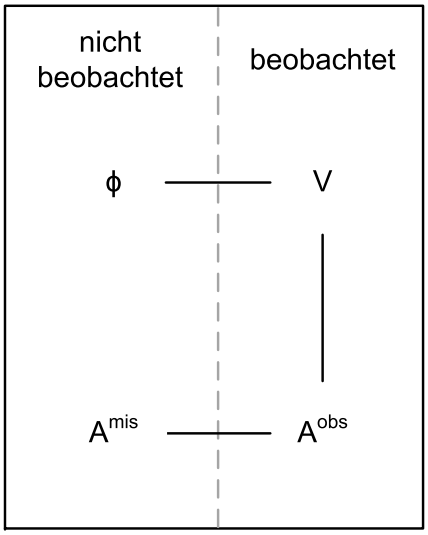


Abbildung 2.9: MAR-Beziehungen im Missing-Data-Modell (in Anlehnung an Schafer und Graham, 2002, S. 152)

Beispiel 2.8: *MAR-Ausfall bei einem Merkmal*
Gegeben sei ein Satz vollständiger Merkmalsvektoren A^{cv} und ein unvollständiger Merkmalsvektor a_{-1}^{mv}. So liegt ein MAR-Ausfallmechanismus vor, wenn:

$$Pr\left(v_{i1}^{mv} = 1|a_{i-}^{cv};a_{i1}^{mv};\phi\right) = Pr\left(v_{ik} = 1|a_{i-}^{cv};\phi\right). \tag{2.9}$$

Die Ausfallwahrscheinlichkeit des Wertes von a_{i1}^{mv} hängt lediglich von den beobachteten Werten a_{i-}^{cv} des Objekts i und einem weiteren unbekannten Parameter ϕ ab.

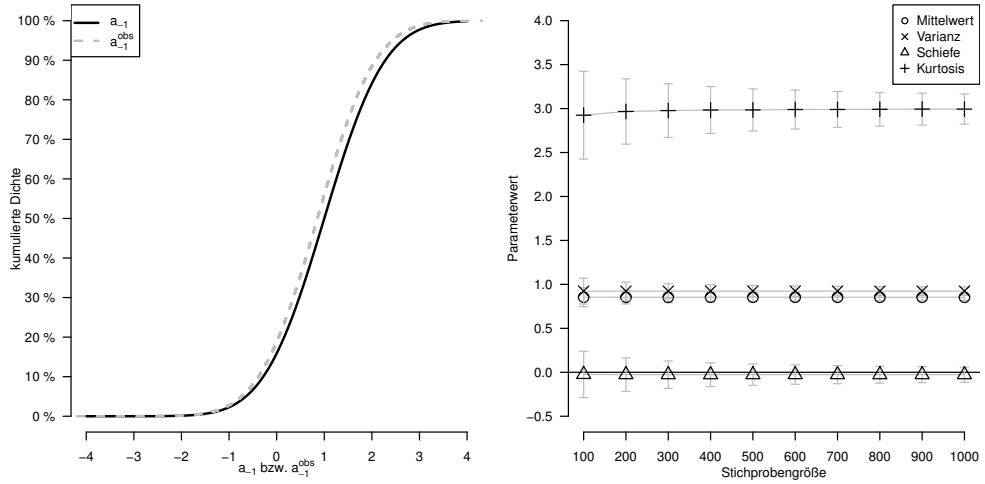
Die MAR-Annahme ist also weniger restriktiv als die MCAR-Annahme, da nun angenommen wird, dass es einen Grund für das Fehlen der Werte gibt. Die A^{obs} stellen somit keine Stichprobe von A dar. Dieses hat zur Konsequenz, dass bei einer Analyse nur von den vorhandenen Daten mit verzerrten Ergebnissen zu rechnen ist. Hierbei hängt die Stärke der zu erwartenden Verzerrungen maßgeblich von der Stärke des Zusammenhangs zwischen A^{obs} und A^{mis} ab (vgl. Beispiel 2.9). Da aber die Gründe für den Ausfall in A^{obs} vorhanden sind, können und, im Gegensatz zum MCAR-Fall, müssen die fehlenden Werte prognostiziert werden, um korrekte Aussagen über A treffen zu können.

Beispiel 2.9: *Auswirkungen von MAR-Ausfall auf eine Normalverteilung*
Gegeben seien zwei normal verteilte Merkmalsvektoren a_{-1} und a_{-2} mit $\mu = (1;1)^T$, $\sigma_{11} = \sigma_{22} = 1$ und $\sigma_{12} = \sigma_{12} = 0{,}5$ und folgender MAR-Ausfallmechanismus:

$$Pr\left(v_{i1} = 1 | a_{i2}\right) = \frac{\exp(-6 + 3 \cdot a_{i2})}{1 + \exp(-6 + 3 \cdot a_{i2})} \; \forall i = 1, \ldots, n. \tag{2.10}$$

So beträgt $\tilde{v}^{mis} = \tilde{v}^{mis}_{\bullet 1}$, unabhängig von σ_{12}, ungefähr 19,4%. Deutlich zu erkennen ist, dass sich die Verteilungen von a^{obs}_{-1} und a_{-1} unterscheiden (vgl. Abbildung 2.10a). Der Ausfallmechanismus (2.10) führt dazu, dass die Ausfallwahrscheinlichkeit mit zunehmendem Wert von a_{-1} steigt und somit die Verteilung von a^{obs}_{-1} links schief ist. Mittelwert, Varianz und Schiefe von a_{-1} können bei MAR, im Gegensatz zu dem MCAR-Beispiel 2.6, nicht erwartungstreu aus a^{obs}_{-1} geschätzt werden[9] (vgl. Abbildung 2.10b). Der Mittelwert und die Varianz von a^{obs}_{-1} sind deutlich kleiner als jene, die für a_{-1} zu erwarten sind. Für eine hinreichend unverzerrte Schätzung der Parameter von a_{-1} müssen zusätzlich zu a^{obs}_{-1} weitere Informationen, wie etwa a_{-2}, hinzugezogen werden.

[9] Ergebnisse der Auswertung von 100.000 simulierten Datensätzen pro Stichprobengröße.



(a) Verteilungsfunktionen (b) Parameterschätzungen

Abbildung 2.10: Vergleich von a_{-1} und a_{-1}^{obs} bei MAR-fehlenden Daten

Der in Formel (2.8) dargestellte Sachverhalt ist eine allgemeingültige Darstellung der MAR-Annahme. Sie gilt insbesondere in dieser Form, wenn die einzelnen Ausprägungen des von Ausfall betroffenen Merkmals nicht voneinander unabhängig sind, wie es bei multivariaten Zeitreihen der Fall ist. Sind die Objekte in A unabhängig voneinander[10], wie bei einfachen Querschnittsstudien angenommen werden kann (vgl. Schnell, 1986, S. 94), lässt sich die Formel (2.8) vereinfachen. Kann diese Annahme getroffen werden, so werden Ausfallabhängigkeiten innerhalb einzelner Variablen ausgeschlossen. V^{mv}, die Verteilung von V in den teilweise beobachteten Merkmalsvektoren, kann nur noch von Ausprägungen in den vollständig beobachteten Merkmalen abhängen. Die Vereinfachung von Formel (2.8) lautet in diesem Fall wie folgt:

$$Pr\left(v_{ij}^{mv} = 1|A;\phi\right) = Pr\left(v_{ij}^{mv} = 1|a_{i-}^{cv};\phi\right) \;\forall\, i = 1,\ldots,r; j = 1,\ldots,q. \quad (2.11)$$

Die Verteilung von V in den verbleibenden, vollständig vorhandenen Merkmalsvektoren V^{cv} entspricht einer Einpunktverteilung mit

$$Pr\left(v_{ij}^{cv} = 0\right) = 1 \;\forall\, i = 1,\ldots,n; j = 1,\ldots,m-q.$$

[10] Dies ist beispielsweise der Fall, wenn A eine einfache Stichprobe (vgl. Pokropp, 1996, S. 27 f.) ist.

Beispiel 2.10: *Grahams Beispiel zum MAR-Ausfall*
Graham (2012, S. 13) bringt ein klassisches Beispiel für einen MAR-Ausfallmechanismus bei langen Fragebögen, bei denen es eine Zeitbegrenzung für die Beantwortung der Fragen gibt. Existiert eine Zeitbegrenzung (beispielsweise 50 Minuten) zur Beantwortung langer Fragebögen, so ist die Ausfallwahrscheinlichkeit der später im Fragebogen auftretenden Fragen maßgeblich von der Lesegeschwindigkeit der Probanden abhängig. Schnelle Leser werden die Befragung in der dafür vorgesehenen Zeit beenden können. Langsame Leser werden dies nicht schaffen. Die Lesefertigkeit der Probanden ist jedoch etwas, das relativ leicht zu Beginn des Fragebogens gemessen werden kann, so dass Daten hierzu für alle Probanden vorhanden sein sollten. Des Weiteren besteht eine deutliche Abhängigkeit der Ausfallwahrscheinlichkeit von der Nummerierung der Fragen. Durch eine Inklusion dieser Variablen im Analysemodell können Verzerrungen, die durch die Lesefertigkeit der Probanden entstehen würden, behoben werden (vgl. Graham, 2009, S. 13).

2.3.3 Not Missing at Random (NMAR)

Ein der Bezeichnung nach relativ neuer Ausfallmechanismus ist not missing at random oder auch kurz NMAR (vgl. Little und Rubin, 2002, S. 12 oder Dillman et al., 2002, S. 18). Ursprünglich von Rubin (1976a) nicht explizit benannt, ist dieser Mechanismus zudem unter der Bezeichnung missing not at random (kurz MNAR) in der Literatur zu finden (vgl. Schafer und Graham, 2002; Graham, 2009; Enders, 2010). Die Bezeichnung NMAR trifft auf die fehlenden Werte einer Datenmatrix zu, wenn die Bedingung

$$f\left(V|A;\phi\right) = f\left(V|A^{obs};A^{mis};\phi\right) \;\forall\, A^{obs}, A^{mis}, \phi \tag{2.12}$$

erfüllt ist. Verbal bedeutet dies, dass nun die Verteilung der fehlenden Werte nicht nur von beobachteten Werten abhängen kann. Vielmehr wird unter dieser Annahme auch zugelassen, dass der Datenausfall nun auch durch die Werte, die beobachtet wären, würden sie nicht fehlen, bedingt sein kann. Dies ist der Fall, wenn entweder eine vollständig unbeobachtete Variable für den Ausfall in einer erhobenen verantwortlich ist oder ein Bezug zwischen den Werten einer Variable und ihrer Ausfallwahrscheinlichkeiten besteht. Die durch die Formel

(2.12) gegebenen Beziehungen können zusätzlich Abbildung 2.11 entnommen werden.

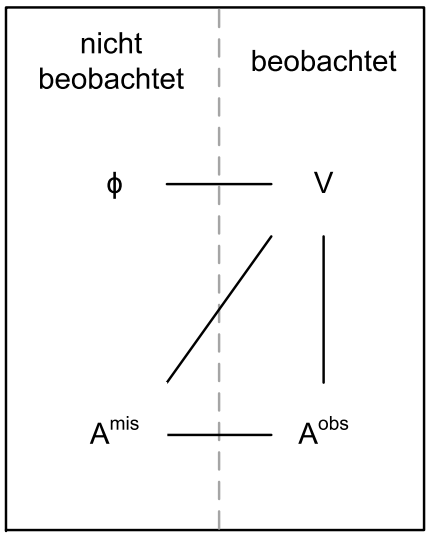


Abbildung 2.11: NMAR-Beziehungen im Missing-Data-Modell (in Anlehnung an Schafer und Graham, 2002, S. 152)

Beispiel 2.11: *NMAR-Ausfall bei einem Merkmal*
Gegeben sei ein Satz vollständiger Merkmalsvektoren A^{cv}, ein nicht beobachtetes Merkmal a_{-1}^{mv} und ein unvollständiges Merkmal a_{-2}^{mv}. So liegt ein NMAR-Ausfallmechanismus vor, wenn:

$$Pr\left(v_{i2}^{mv} = 1 | a_{i-}^{cv}; a_{i-}^{mv}; \phi\right) = Pr\left(v_{i2}^{mv} = 1 | a_{i1}^{mv}; \phi\right) \ \forall\, i = 1, \ldots, n. \tag{2.13}$$

Die Ausfallwahrscheinlichkeit der Werte in Variable a_{-2}^{mv} hängt also lediglich von dem Wert einer nicht beobachteten Variable a_{-1}^{mv} und einem weiteren unbekannten Parameter ϕ ab.

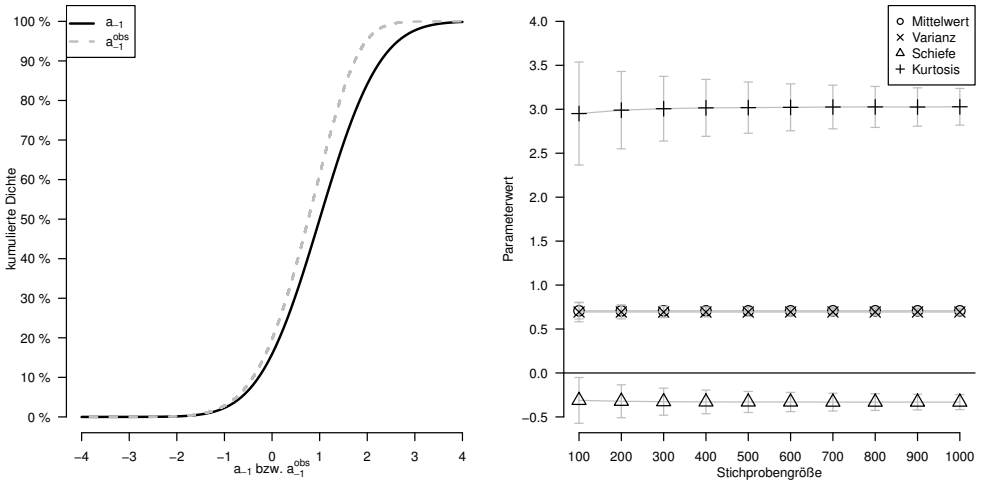
NMAR ist die letzte Klasse an Ausfallmechanismen, deren Erkennung relevant für die weitere Behandlung von fehlenden Werten ist. Wie bereits der Bezeichnung zu entnehmen ist, sind dies jegliche Ausfallmechanismen, die nicht der MAR-Annahme entsprechen. Es handelt sich somit bei Daten, die nicht der MAR-Bedingung genügen, um systematische Ausfallmechanismen (vgl. Bankhofer, 1995, S. 22). Somit stellt A^{obs} unter der NMAR-Annahme keine Stichprobe von A dar, und da V von A^{mis} abhängt, können die fehlenden Werte nicht aus den vorhandenen prognostiziert werden. Gemäß Little und Rubin

(2002, S. 13) führt eine Analyse, basierend auf A^{obs}, bei Vorliegen eines NMAR-Ausfallmechanismus im Allgemeinen zu verzerrten Ergebnissen. Ein korrekter Umgang mit den fehlenden Werten erfordert also an dieser Stelle zusätzliche Informationen[11] beziehungsweise Annahmen[12].

Beispiel 2.12: *Auswirkungen von NMAR-Ausfall auf eine Normalverteilung*
Gegeben sei ein normal verteilter Merkmalsvektor a_{-1} mit $\mu = (1)$ und $\sigma_{11} = 1$, der von dem folgenden NMAR-Ausfallmechanismus betroffen ist:

$$Pr\left(v_{i1} = 1 | a_{i1}\right) = \frac{\exp(-6 + 3 \cdot a_{i1})}{1 + \exp(-6 + 3 \cdot a_{i1})} \; \forall i = 1, \ldots, n. \tag{2.14}$$

So beträgt $\tilde{v}^{mis} = \tilde{v}^{mis}_{\bullet 1}$, abhängig vom Wert der Variable a_{-1}, circa 19,4%. Deutlich zu erkennen ist, dass sich die Verteilungen von a^{obs}_{-1} und a_{-1} unterscheiden (vgl. Abbildung 2.12a). Die Verteilung von a^{obs}_{-1} ist auffallend links schief ($\gamma_1 \approx -0{,}33$), welches darauf zurückzuführen ist, dass kleine Werte von a_{-1} deutlich seltener fehlen als große Werte.



(a) Verteilungsfunktionen (b) Parameterschätzungen

Abbildung 2.12: Vergleich von a_{-1} und a^{obs}_{-1} bei NMAR-fehlenden Daten

Auch der Erwartungswert und die Varianz von a^{obs}_{-1} sind mit $\mu_1 \approx 0{,}707$ und

[11] Diese können beispielsweise auch durch Folgebefragungen erhoben werden (Glynn et al., 1993; Graham und Donaldson, 1993).

[12] Von primärer Bedeutung sind hier die Annahmen, die für eine Modellierung des Ausfallmechanismus oder der von Ausfall betroffenen Variable erforderlich sind (vgl. Little und Rubin, 2002, S. 312 ff.).

$\sigma_{11} \approx 0{,}693$ deutlich kleiner als bei a_{-1} (vgl. Abbildung 2.12b)[13]. Daher ist deutlich, dass die Parameter der Verteilung von a_{-1} bei NMAR, im Gegensatz zu MCAR (vgl. Beispiel 2.6), nicht erwartungstreu aus a_{-1}^{obs} geschätzt werden können. Für eine hinreichend unverzerrte Schätzung der Parameter von a_{-1} müssen zusätzlich zu a_{-1}^{obs} weitere Informationen hinzugezogen werden.

Ähnlich wie im Abschnitt 2.3.2 für die MAR-Annahme dargestellt, handelt es sich bei Formel (2.12) um die allgemeingültige Darstellung des NMAR-Ausfallmechanismus. Wird auch hier die sinnvolle Annahme, dass alle Objekte unabhängig voneinander sind, getroffen, so lässt sich auch die Darstellung des NMAR-Ausfallmechanismus vereinfachen. Unter Unabhängigkeit der Objekte untereinander trifft die Bezeichnung NMAR auf fehlende Werte einer Datenmatrix zu, sofern die Bedingung

$$Pr\left(v_{ij}^{mv} = 1|A;\phi\right) = Pr\left(v_{ij}^{mv} = 1|a_{i-};\phi\right) \;\forall\, i = 1,\ldots,r; j = 1,\ \ldots, q \quad (2.15)$$

erfüllt ist, und die Verteilung von V^{cv} folgender Einpunktverteilung entspricht:

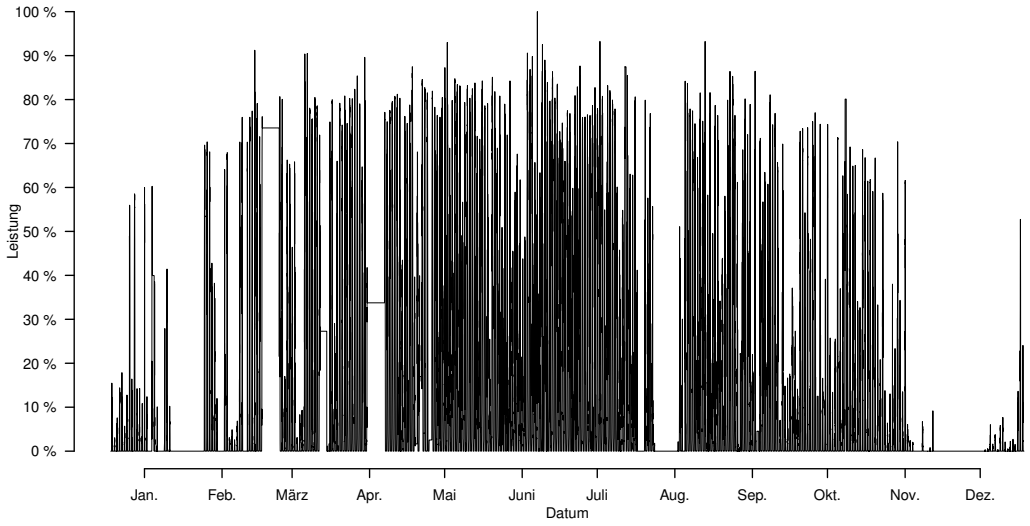
$$Pr\left(v_{ij}^{cv} = 1\right) = 0 \;\forall\, i = 1,\ldots,n; j = 1,\ldots,m-q.$$

Beispiel 2.13: *NMAR-Ausfall bei einer Zeitreihe*
Die in der folgenden Grafik 2.13 dargestellte Zeitreihe beinhaltet Daten eines Messsensors einer Photovoltaik-Anlage des Fraunhofer-Instituts für Optronik, Systemtechnik und Bildauswertung in Ilmenau, Deutschland. Messwerte der Ausgangsleistung des Jahres 2010 wurden auf zwischen $[0;1]$ normiert. Da es sich um eine stationäre Anlage ohne Nachführung in Neigung und Ausrichtung handelt, ergeben sich klar erkennbare Saisonkomponenten im Kurvenverlauf. Deutlich in Abbildung 2.13 zu erkennen sind die Tages- und Monatskomponenten. Zudem sind einige Bereiche mit fehlerhaften Werten, deren Entfernung zu fehlenden Werten führt, ersichtlich.

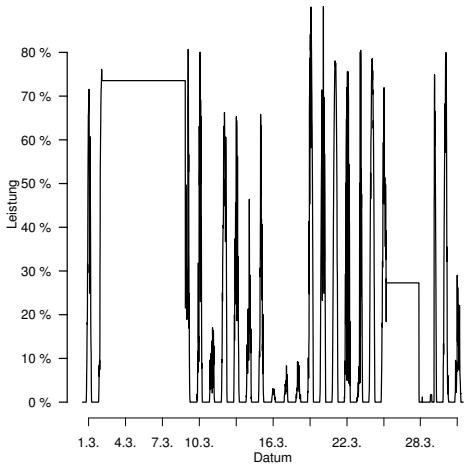
Tageszeiten, bei denen keine (beispielsweise im Zeitraum vom 7.8. bis 14.8. in der Abbildung 2.13c) oder über längere Zeiträume dieselbe Ausgangsleistung verzeichnet wurde (beispielsweise im Zeitraum vom 2.3. bis 9.3. in der Abbildung 2.13b), sind fehlerhaft und müssen eliminiert werden. Die Gründe für den sich hierdurch ergebenden Datenausfall sind mannigfaltig und nur zum

[13] Ergebnisse der Auswertung von 100.000 simulierten Datensätzen pro Stichprobengröße.

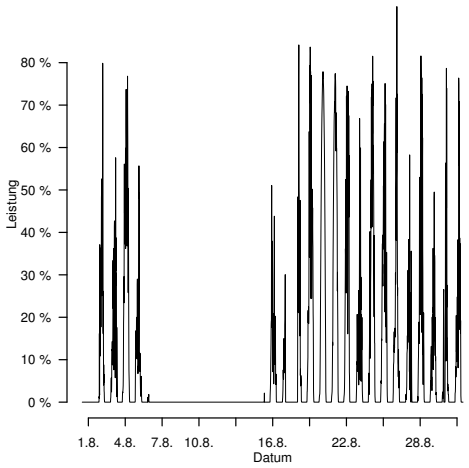
Teil nachträglich zu erfassen. Häufig sind technische Defekte für den Ausfall von richtigen Messwerten verantwortlich, die, mangels einer täglichen Überwachung der Datenerfassung, zu spät erkannt werden und sich deswegen auch über längere Zeiträume erstrecken. Die langen Messwertausfälle im November und Februar können auf eine Wechselwirkung zwischen Niederschlag und Temperatur (Schneefall) zurückgeführt werden. Da jedoch weder die klimatischen Verhältnisse noch andere Merkmale, die einen Bezug zur Ausfallursache aufweisen, zusätzlich erfasst werden, sind die existierenden Ausfallmechanismen NMAR. Zum Umgang mit den fehlenden Werten sind also zusätzliche Informationen in Form von Daten oder Expertenwissen notwendig.



(a) Leistungsverlauf Jahr 2010



(b) Leistungsverlauf März 2010



(c) Leistungsverlauf August 2010

Abbildung 2.13: Ausgangsleistung einer Photovoltaik-Anlage im Jahr 2010

2.3.4 Plausibilität des NMAR-Mechanismus

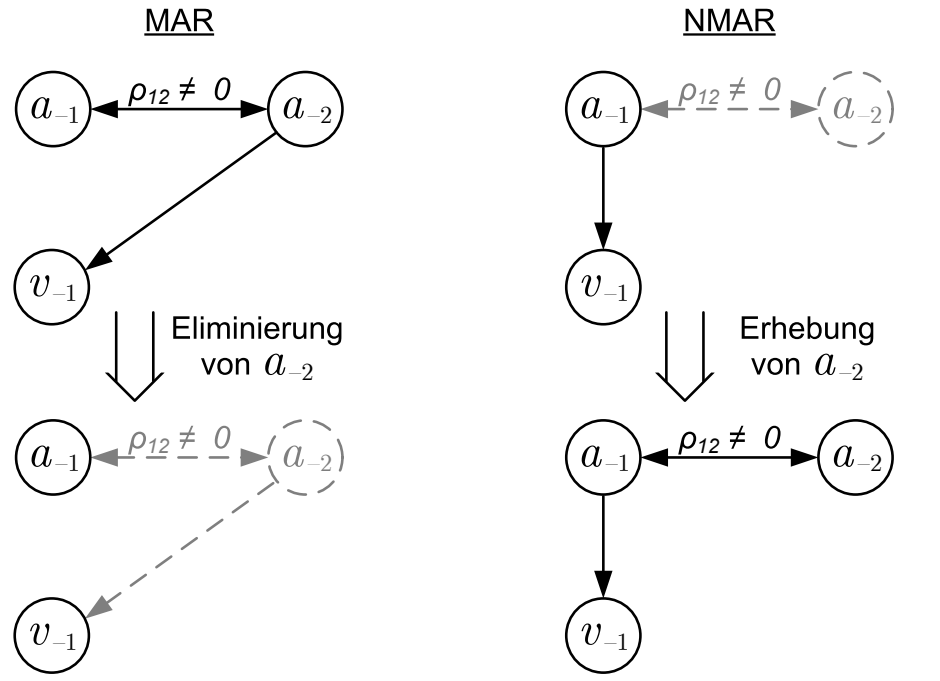
Da grundsätzlich nur statistische Tests bezüglich des Zutreffens der MCAR-Annahme existieren (vgl. Bankhofer, 1995, S. 76 beziehungsweise Enders, 2010, S. 17), ist die Frage, insbesondere für den Praktiker, wie zwischen MAR und NMAR zu differenzieren ist, von übergeordneter Bedeutung. Das Vorliegen von NMAR führt zu deutlich größeren Problemen als das Vorliegen von MAR und mündet im Extremfall in die Entscheidung, dass die Daten nicht analysiert werden können und neu erhoben werden müssen. Daher besteht ein deutliches Interesse, das Vorliegen eines MAR-Mechanismus zu plausibilisieren und sogleich Gegenargumente für das Vorliegen eines NMAR-Mechanismus zu finden.

Ist das Vorliegen eines NMAR-Mechanismus in Daten, wie sie in den Sozialwissenschaften typischerweise vorzufinden sind, plausibel? Zur Beantwortung dieser Frage wird zunächst ein kontrafaktisches Gedankenexperiment eingesetzt. In einem solchen werden nicht reale, praktisch meist nicht realisierbare, Situationen vorgestellt und zum Zwecke des Erkenntnisgewinns gedanklich manipuliert[14]. Das hiesige Experiment soll durch die zwei in Abbildung 2.14 dargestellten Beispiele unterstützt werden. Der obere Teil von Abbildung 2.14a ist eine qualitative Darstellung des MAR-Ausfallmechanismus, der bereits im Beispiel 2.9 präsentiert wurde, während der NMAR-Ausfallmechanismus aus Beispiel 2.12 qualitativ im oberen Teil von Abbildung 2.14b dargestellt ist.

Wird bei dem MAR-Ausfallmechanismus (in der Abbildung 2.14a oben) die für das Fehlen verantwortliche Variable a_{-2} entfernt, entsteht jener NMAR-Mechanismus, der in der Abbildung unten dargestellt wird. Dieser Ausfallmechanismus kann, rein auf Grund des vorliegenden Datenmaterials, nur nach der Intensität der Wirkung von dem NMAR-Ausfallmechanismus (in der Abbildung 2.14b oben dargestellt) unterschieden werden. Betrachtet man nun die Kehrseite und erhebt bei jenem in Abbildung 2.14b oben dargestelltem Fall zusätzlich die Variable a_{-2}, stellt sich die Frage, ob nun weiterhin ein NMAR-Mechanismus vorliegt. Von einer rein theoretischen Betrachtungsweise her muss das Vorliegen eines NMAR-Ausfallmechanismus bejaht werden. Von einer praktischen Betrachtungsweise her ist dies aber fraglich. Ein NMAR-

[14] Der Einsatz von Gedankenexperimenten erfreut sich nicht nur in den Sozialwissenschaften großer Beliebtheit. Zum Nutzen und zur Verwendung von Gedankenexperimenten sei an dieser Stelle auf Popper (1968, S. 464 ff.) verwiesen.

Ausfallmechanismus ist in diesem Fall durch den vorhandenen Zusammenhang – in diesem Fall die Korrelation ρ_{12} – zwischen den Variablen a_{-2} und a_{-1} bedingt. Ist der Zusammenhang zwischen den beiden Variablen „stark“, lässt sich kaum im Ergebnis zwischen den NMAR- und MAR-Mechanismen unterscheiden. Eine Trennung zwischen den beiden Ausfallmechanismen anhand ihrer Auswirkungen ist deutlich schärfer, wenn der Zusammenhang zwischen den beiden Variablen „niedrig“ oder gar nicht vorhanden ist.



(a) MAR-Ausfallmechanismus (b) NMAR-Ausfallmechanismus

Abbildung 2.14: Beispielausfallmechanismen des Gedankenexperiments

Zudem ist es möglich, bei jedem vorliegenden Datenmaterial zwei Ausfallmechanismen, einen MAR und einen NMAR, zu modellieren, die eine äquivalente Verteilung der beobachteten Werte erzeugt (vgl. Beispiel 2.14). Beispiel 2.15 verdeutlicht, dass, je nach der Stärke der existierenden Zusammenhänge, ein MAR-Mechanismus dieselbe Wirkung entfalten kann wie ein NMAR-Mechanismus.

Beispiel 2.14: *In der Wirkung äquivalenter MAR und NMAR-Ausfall*
Gegeben seien zwei normal verteilte Variablen a_{-1} und a_{-2} mit $\mu = (1;1)^T$,

$\sigma_{11} = \sigma_{22} = 1$ und $\sigma_{12} = 0{,}5$ sowie der folgende NMAR-Ausfallmechanismus:

$$Pr\left(v_{i1} = 1 | a_{i1}\right) = \frac{\exp(-\sqrt{6} + \frac{\sqrt{3}}{2} \cdot a_{i1})}{1 + \exp(-\sqrt{6} + \frac{\sqrt{3}}{2} \cdot a_{i1})} \; \forall i = 1, \ldots, n. \tag{2.16}$$

Dieser Ausfallmechanismus ist von der Wirkung auf a_{-1} nicht von jenem aus Beispiel 2.9 (Formel (2.10)) zu unterscheiden. Beide Ausfallmechanismen erzeugen Variablen a_{-1}^{obs} mit derselben Verteilung, was auch analytisch oder mittels Simulation gezeigt werden kann. Der Wirkungszusammenhang der Ausfallmechanismen wird in Abbildung 2.15 zusammenfassend dargestellt.

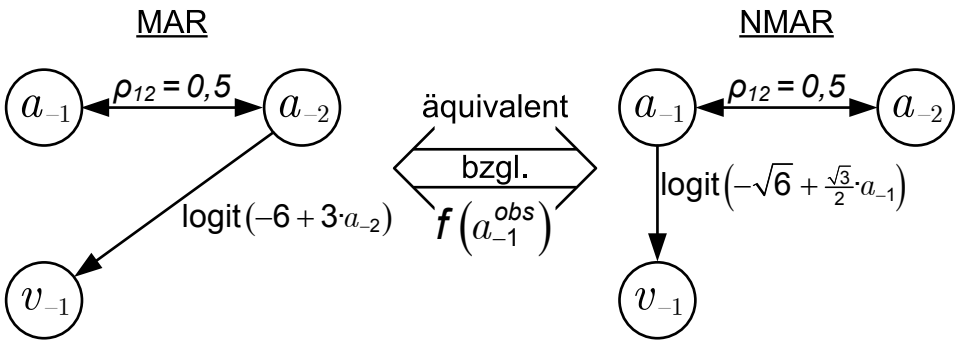
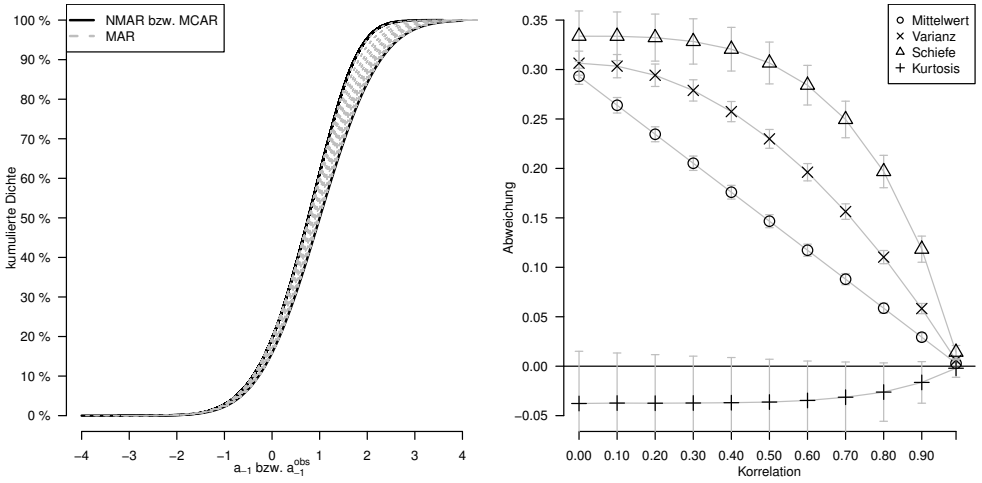


Abbildung 2.15: Zwei in der Wirkung äquivalente Ausfallmechanismen

Beispiel 2.15: *Auswirkungen eines MAR-Ausfalls in Abhängigkeit von ρ*
Gegeben seien zwei normal verteilte Merkmalsvektoren a_{-1} und a_{-2} mit $\mu = (1; 1)^T$, $\sigma_{11} = \sigma_{22} = 1$ und $\sigma_{12} = \sigma_{12}$ sowie die Ausfallmechanismen gemäß der Formeln (2.10) und (2.14) der Beispiele 2.9 und 2.12. Wird nun die Korrelation ρ_{12} im Bereich $[0; 1]$ variiert, lassen sich unterschiedliche Verteilungen für a_{-1}^{obs} im MAR-Fall ermitteln[15] (vgl. Abbildung 2.16a). Deutlich zu erkennen ist, dass bei $\rho_{12} = 0$ die Verteilung von a_{-1}^{obs} der Verteilung von a_{-1} entspricht und somit der Wirkung eines MCAR-Ausfalls entspricht. Bei einer Korrelation von $\rho_{12} = 1$ entspricht die Verteilung von a_{-1}^{obs} der Verteilung von a_{-1}^{obs} unter dem gewählten NMAR-Ausfallmechanismus. Dass dieser MAR-Mechanismus mit steigender Korrelation der zwei Variablen zu einem NMAR-Mechanismus konvergiert, zeigt zudem Abbildung 2.16b. Diese mittels Simulation berechne-

[15] Ergebnisse der Auswertung von 100.000 simulierten Datensätzen mit einer Stichprobengröße von 10.000 pro Korrelation. Als Korrelation wurde $\rho_{12} = (0{,}0(0{,}1)0{,}9; 0{,}99)$ gewählt.

ten Werte belegen, dass die Unterschiede in den Parametern von a_{-1}^{obs} zwischen den Ausfallmechanismen mit steigendem ρ_{12} verschwinden.



(a) Verteilungsfunktionen (b) Parameterschätzungen

Abbildung 2.16: Vergleich von a_{-1}^{obs} bei MAR- und NMAR-fehlenden Daten

Da eine ähnliche Argumentationslinie auch für andere Formen von NMAR gezogen werden kann (vgl. Enders, 2010, S. 14–17), wird in der Literatur eine inklusive Analysestrategie empfohlen (Rubin, 1996; Schafer, 1997, S. 22-23; Collins et al., 2001; Schafer und Graham, 2002). Eine inklusive Analysestrategie bedeutet, dass eine Menge an Auxiliarvariablen mit in die Analyse oder den Imputationsalgorithmus einbezogen wird. Diese zusätzlichen Variablen werden auxiliar genannt, weil ihre Analyse nicht von primärem Interesse ist. Sie werden lediglich verwendet, um bessere Ergebnisse zu erhalten. Zwei Arten von Hilfsvariablen können von Interesse sein (vgl. Collins et al., 2001, S. 331). Bei der ersten Art besteht ein Zusammenhang zwischen der Auxiliarvariable und dem Ausfallmechanismus. Die zweite Art von Hilfsvariable ist eine Kovariate jener Variable, die fehlende Werte aufweist. Größtenteils bleibt aber offen, wie stark die Zusammenhänge sein müssen, damit etwaige Verzerrungen substantiell reduziert werden können. Lediglich die Untersuchungen von Collins et al. (2001, S. 345) untermauern die obigen Argumentationen mit einer Simulationsstudie. Mittels dieser wurde gezeigt, dass die inklusive Strategie zu substantiell verbesserten Ergebnissen bei NMAR-fehlenden Daten führen kann[16].

[16] Es handelt sich hierbei um „Study 3“ des Aufsatzes.

Aufgrund der bisherigen Erläuterungen, sowie der in der Literatur vorhandenen Diskussionen und empirischen Untersuchungen, ist die Antwort auf die Frage, ob ein NMAR-Ausfallmechanismus plausibel ist, klar. Die Plausibilität dessen, ob ein NMAR-Ausfallmechanismus vorhanden ist, richtet sich nach der Frage ob vollständig beobachtete Merkmale vorhanden sind, welche einen hinreichend großen Erklärungswert für das unvollständige Merkmal aufweisen. Folglich ist die Präsenz, und die hiermit auftretenden Probleme, eines NMAR-Ausfallmechanismus nur plausibel, wenn alle erhobenen Merkmale gemeinsam keinen hinreichenden Erklärungswert für jedes unvollständige Merkmal aufweisen. Hieraus ergeben sich jene für den Umgang mit fehlenden Werten wohl interessantesten Fragen:

1. Welcher Erklärungswert, den die vorhandenen Merkmale für die unvollständigen Merkmale aufweisen, ist hinreichend, um einen MAR-Ausfallmechanismus annehmen zu können?

2. Welche Methode zum Umgang mit fehlenden Werten nutzt den vorhandenen Erklärungswert am besten?

Während die zweite Frage in der Missing-Data-Literatur eine prominente Rolle einnimmt, und hier immer noch eine Weiter- und Neuentwicklung von Methoden stattfindet, wurde die erste Frage bis heute nicht zufriedenstellend beantwortet.

2.4 Missing-Data-Methoden

Nachdem die Frage nach der Art des Ausfallmechanismus beantwortet wurde, stellt sich unweigerlich die Frage danach, wie mit den fehlenden Werten verfahren werden soll. Hierzu wurde in der Literatur über Jahrzehnte hinweg eine Vielzahl an Methoden entwickelt und zur Anwendung, in Abhängigkeit des vorliegenden Ausfallmechanismus, vorgeschlagen. Mit der Häufung an entwickelten Methoden entstanden Taxonomien, welche die entwickelten Methoden nach ihren grundlegenden Eigenschaften systematisierten. Die Bandbreite dieser Klassifikationsschemata reicht von simpel bis zu sehr komplex.

Einige Autoren beschränken ihre Systematisierung der Verfahren auf zwei Klassen. Beispielsweise unterscheiden Schafer und Graham (2002, S. 155 ff.)

lediglich zwischen der Maximum-Likelihood-Schätzung und den sogenannten „Older Methods“, worunter alle verbleibenden Verfahren zusammengefasst sind. Auch Peughd und Enders (2004, S. 528) differenzieren zwischen den „traditionellen Verfahren“ und „modernen Techniken.“ In ähnlicher Form unterteilt Roth (1994, S. 539 ff.) seine Darstellung von den Missing-Data-Methoden in „simple Techniken“ und „Maximum Likelihood und verwandte Methoden“. Derartige Aufteilungen erscheinen für eine Darstellung von Missing-Data-Methoden wenig zweckmäßig. Diese Aufteilungen bringen teilweise Wertungen in die Klassifikation mit ein oder suggerieren, dass gewisse Methoden keiner weiteren Entwicklung oder Nutzung unterliegen.

Demgegenüber stehen Systematisierungen in der Literatur, welche eher inhaltlich an den Methoden ausgerichtet sind. Einige Autoren beschränken sich in ihrer Darstellung der Methoden zum Umgang mit fehlenden Daten rein auf Imputations- beziehungsweise auch Gewichtungsmethoden (vgl. Kalton und Kasprzyk, 1986, S. 1). Eine deutlich umfassendere Klassifikation wird von Bankhofer (1995, S. 89) vorgenommen. Hier werden die Missing-Data-Methoden in Eliminierungs-, Imputations-, Parameterschätz- und multivariate Analyseverfahren sowie Sensitivitätsbetrachtungen unterteilt. Little und Rubin (2002, S. 19 f.) unterscheiden hingegen zwischen Methoden, die auf vollständigen Objekten basieren, Gewichtungs- und Imputationsverfahren sowie modellbasierten Verfahren.

Im Folgenden werden einige Methoden zum Umgang mit fehlenden Daten im Überblick dargestellt. Beschränkt wird sich auf Eliminierungs- und Imputationsverfahren, welche, gemäß den Untersuchungen von Bodner (2006, S. 677) sowie Peughd und Enders (2004, S. 541), die am häufigsten angewendeten Methoden sind. Hierbei befasst sich Abschnitt 2.4.1 mit Verfahren, die Objekte oder Merkmale mit fehlenden Daten systematisch von der Analyse ausschließen. Abschnitt 2.4.2 behandelt Verfahren, die zur Ersetzung von fehlenden mit plausiblen Werten verwendet werden können. Darstellungen von weiteren Verfahrensmöglichkeiten sind für Eliminierungs-, Imputations- und multivariate Analyseverfahren beispielsweise Bankhofer (1995) zu entnehmen. Umfassende Darstellungen der Parameterschätzverfahren, insbesondere Beschreibungen des EM-Algorithmus (Dempster et al., 1977), sind in den Arbeiten von Little und Rubin (2002), Enders (2010), Allison (2001) und Schafer (1997) enthalten.

2.4.1 Eliminierungsverfahren

Unter den Eliminierungsverfahren sind alle Methoden zur Behandlung von fehlenden Daten zu subsumieren, bei denen Objekte oder Merkmale von der Untersuchung auf systematische Art und Weise ausgeschlossen werden. Diese Verfahren sind, insbesondere in den Sozialwissenschaften, die am häufigsten angewandten Methoden zum Umgang mit fehlenden Daten (vgl. Enders, 2010, S. 39). Belegt wird dieses unter anderem durch die Studien von Bodner (2006) und Schlomer et al. (2010) im Bereich Psychologie, Peughd und Enders (2004) im Bereich Bildungsforschung, Wood et al. (2004) im Bereich Medizin und Backhaus und Blechschmidt (2009) im Bereich Betriebswirtschaftslehre. Trotz ihrer Beliebtheit lassen diese Verfahren jedoch lediglich unter der restriktiven Annahme des MCAR-Ausfallmechanismus, das heißt die vorhandenen Daten stellen eine Stichprobe der gewünschten Daten dar, korrekte Inferenzen zu (vgl. Little und Rubin, 2002, S. 41–42). Jedoch selbst in der selten zu erwartenden Situation, dass die Daten der MCAR-Annahme genügen, führt die Anwendung von Eliminierungsverfahren im besten Fall zu einem Verlust an Präzision. In jeder anderen Situation werden die Daten mittels der Verfahren verzerrt.

Grundsätzlich existieren die in Abbildung 2.17 dargestellten vier Eliminierungsverfahren. Wie aus der Grafik ersichtlich, können die Verfahren zunächst danach differenziert werden, ob Objekte oder Merkmale von den Analysen ausgeschlossen werden sollen. Des Weiteren kann unterschieden werden, ob eine Eliminierung vollständig a priori erfolgt oder ob die Eliminierung ad-hoc auf jene Objekte oder Merkmale angewandt wird, die in einer konkreten Analyse verwendet werden sollen. In der Literatur zu fehlenden Daten erfolgt eine Behandlung der fehlenden Werte nahezu ausnahmslos mit dem Ziel, dass in folgenden Analysen Aussagen über die erhobenen Merkmale generiert werden sollen. Vor diesem Hintergrund ist es verständlich, dass kaum Arbeiten existieren, die sich mit der Eliminierung von Merkmalen beschäftigt. Schließlich werden die Merkmale in einer Analyse nicht grundlos erhoben, sondern meist sollen die Zusammenhänge zwischen diesen mittels der Objekte – auch Merkmalsträger genannt – betrachtet werden. Es werden daher im Folgenden lediglich die Objekt-Eliminierungsverfahren dargestellt.

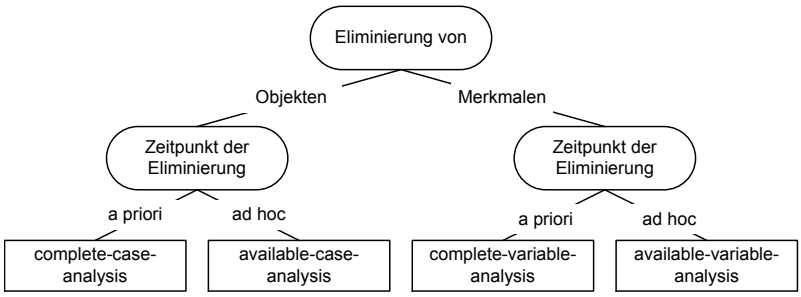


Abbildung 2.17: Die möglichen Eliminierungsverfahren

2.4.1.1 Analyse der vollständigen Objekte

Die Analyse vollständig vorhandener Objekte (complete case analysis oder auch listwise deletion) beschränkt sich in der Auswertung auf jene Objekte, die bezüglich aller erhobenen Merkmale vollständig vorhanden sind (vgl. Bankhofer, 1995, S. 91; Little und Rubin, 2002, S. 41; Enders, 2010, S. 39). Für die Complete-Case-Analyse wird lediglich eine kleinere $(n - r \times m)$ Teilmatrix untersucht. Diese ergibt sich, wenn A so sortiert wird, das erst alle vollständigen Objekte und dann alle Objekte mit mindestens einem fehlenden Wert aufgeführt werden (vgl. Gleichung (2.17)).

$$A = \begin{pmatrix} A^{cc} \\ A^{mc} \end{pmatrix} = \begin{pmatrix} (a_{ik}^{cc})_{n-r,m} \\ (a_{ik}^{mc})_{r,m} \end{pmatrix} \tag{2.17}$$

Beispiel 2.16: *Eliminierung bei der 2009 ACS Stichprobe*
Wird sich für die Datenmatrix im Anhang A (Fall 1) auf eine Analyse der vollständig vorhandenen Objekte beschränkt, so werden die Objekte 2, 6, 9, 11, 13, 16, 18, und 23 eliminiert. Folgende Analysen beschränken sich auf die Datenmatrix A^{cc} mit den Dimensionen (17×11).

Vorteile dieser Vorgehensweise sind im Wesentlichen die Einfachheit, mit der sie durchgeführt werden kann, sowie die Vergleichbarkeit aller auf dieser Datenbasis berechneten deskriptiven und induktiven Statistiken, da die Stichprobengröße konstant bleibt. Allerdings überwiegen die Nachteile, die sich im besten Falle auf Verlust von Präzision bei den auf Basis der verbleibenden Daten berechneten Schätzern beschränken. Im schlimmsten Falle, der sich zwangsweise ergibt, sofern die Daten nicht MCAR genügen, wird ein Bias in die Ergebnisse eingeführt. Wird in der Literatur eine Anwendbarkeit dieser

Methode eingeräumt, wird auch darauf verwiesen, dass der Anteil der fehlenden Werte gering sein sollte. Schwab (1991) bezeichnet eine 5%-Grenze für den Prozentsatz fehlender Daten zur Durchführung einer Analyse der vollständigen Objekte als akzeptabel. Little und Rubin (2002, S. 42) argumentieren jedoch, dass eine solche Grenze schwer zu formulieren ist. Schließlich richtet sich die Größe des Bias und Präzisionsverlustes nicht nur nach der Menge der fehlenden Daten, sondern auch nach dem Muster des Fehlens, den zu schätzenden Parametern und inwieweit sich vorhandene und fehlende Daten unterscheiden. Insbesondere für den ungünstigen Fall eines allgemeinen Ausfallmusters lassen sich plakative Beispiele konstruieren.

Beispiel 2.17: *Eliminierung bei MCAR-Ausfall und allgemeinen Muster*
Ist der Ausfallmechanismus MCAR und sind alle Merkmale gleichmäßig im Sinne eines allgemeinen Ausfallmusters betroffen, so ist das Auftreten von fehlenden Werten in einem Merkmal statistisch unabhängig vom Auftreten in jedem anderen. Insofern kann der zu erwartende Anteil von Objekten, die mindestens einen fehlenden Wert aufweisen, durch die Formel (2.18) beschrieben werden.

$$\frac{r}{n} = 1 - \left(1 - \tilde{v}^{mis}\right)^m \tag{2.18}$$

So müssten, exemplarisch, bei 5% fehlenden Werten sowie zwei Merkmalen 9,75% und bei 10 Merkmalen bereits ca. 40,12% der Objekte eliminiert werden.

2.4.1.2 Analyse der verfügbaren Objekte

In einer Analyse der verfügbaren Objekte (available case analysis, oder auch pairwise deletion) erfolgt keine allgemeine, vorab festgelegte Bereinigung der Datenbasis. Vielmehr werden ad hoc jene Objekte von einer Analyse ausgeschlossen, welche in der für die Analyse erforderlichen Merkmale unvollständig sind (vgl. Bankhofer, 1995, S. 93; Little und Rubin, 2002, S. 53 f.; Enders, 2010, S. 40 f.). Eine solche Vorgehensweise erhält eine größere Anzahl von Objekten für die individuellen Analysen und ist somit deutlich sparsamer mit der Menge an eliminierten Daten. Lediglich im Falle dessen, dass alle Merkmale mit fehlenden Daten gleichzeitig analysiert werden sollen, wird dieselbe Anzahl an Objekten wie bei der Complete-Case-Analyse eliminiert. Insbesondere, wenn univariate Analysen durchgeführt werden sollen, erhält diese Vorgehens-

weise deutlich mehr Objekte als eine direkte Vorab-Löschung aller Objekte mit fehlenden Werten. Auch bei multivariaten Analysen, die Merkmale paarweise betrachten, führt eine Analyse der verfügbaren Objekte zu einem geringeren Verlust an statistischer Güte.

Beispiel 2.18: *Korrelationsschätzung bei Eliminierungsverfahren*
In diesem Beispiel soll die Datenmatrix im Anhang A für den Fall betrachtet werden, dass Fall 3 und Fall 4 gleichzeitig vorhanden sind. Zu berechnen ist eine Korrelationsmatrix unter der Verwendung der kardinalen Merkmale a_{-8} bis a_{-11}. Während bei einer Complete-Case-Analyse die Berechnung aller Kovarianzen auf denselben 11 Objekten (2, 3, 6, 9, 14, 16, 19, 21, 22, 24 und 25) beruht, gestaltet sich die Anzahl der verwendeten Objekte in einer Available-Case-Analyse, wie in der Tabelle 2.1 dargestellt.

	UHRSWORK	INCTOT	FTOTINC	INCSS
UHRSWORK	25	18	17	25
INCTOT		18	11	18
FTOTINC			17	17
INCSS				25

Tabelle 2.1: Anzahl der verfügbaren Objekte für eine Korrelationsberechnung, Anhang A, Fälle 3 und 4 gemeinsam wirkend

Bei Berechnung der Kovarianzen unter der Verwendung der jeweils paarweise verfügbaren Objekte entsteht die folgende Kovarianzmatrix:

$$\Sigma = \begin{pmatrix} 394 & 248.264 & 654.155 & -34.879 \\ 248.264 & 3.277.445.870 & 4.784.087.982 & 270.244.788 \\ 654.155 & 4.784.087.982 & 5.532.581.245 & 231.957.850 \\ -34.879 & 270.244.788 & 231.957.850 & 37.530.900 \end{pmatrix}.$$

Werden nun die Pearson-Korrelationen mittels der Varianzen und Kovarianzen von den beteiligten Merkmale berechnet, ergibt sich die folgende Korrelationsmatrix:

$$\mathcal{P} = \begin{pmatrix} 1 & 0{,}218 & 0{,}442 & -0{,}286 \\ 0{,}218 & 1 & 1{,}123 & 0{,}770 \\ 0{,}442 & 1{,}123 & 1 & 0{,}509 \\ -0{,}286 & 0{,}770 & 0{,}509 & 1 \end{pmatrix}.$$

Bei einer Verwendung aller vollständigen Objekte entstehen bei analoger Vorgehensweise die folgenden Korrelationen, die von dem obigen Ergebnis abweichen:

$$\mathcal{P} = \begin{pmatrix} 1 & 0{,}132 & 0{,}433 & -0{,}225 \\ 0{,}132 & 1 & 0{,}749 & 0{,}835 \\ 0{,}433 & 0{,}749 & 1 & 0{,}525 \\ -0{,}225 & 0{,}835 & 0{,}525 & 1 \end{pmatrix}.$$

Anhand der Ausführungen zu Beispiel 2.18 werden Schwächen des Verfahrens deutlich. Zwar sind bei der Available-Case-Analyse immer mindestens genau so viele Objekte vorhanden wie bei der Complete-Case-Analyse, aber die Berechnungsgrundlage kann sich stets verändern. Am Beispiel der Korrelationsanalyse bedeutet dies, dass die Varianzen der Merkmale unter Umständen mit Hilfe einer unterschiedlichen Anzahl an Objekten berechnet werden. Diese können dann auch von der Anzahl der Objekte abweichen, mittels derer die Kovarianzen berechnet werden. Dies kann, je nach vorliegendem Ausfallmuster, dazu führen, dass Korrelationen berechnet werden, die sich außerhalb des Bereichs $[-1; 1]$ befinden (vgl. Beispiel 2.18, $\rho_{23} > 1$). Allerdings zeigt Matthai (1951, S. 148 f.), dass dies mit einem speziellen Schätzer vermieden werden kann. Werden die Korrelationen direkt berechnet, ohne vorherige Berechnung der auf unterschiedlichen Objektmengen basierenden Kovarianzen, kann dieses Problem umgangen werden.

Weitere Probleme behebt selbst die modifizierte Vorgehensweise von Matthai (1951) nicht, wie beispielsweise Little und Rubin (2002, S. 55) zeigen. Soll eine Kovarianzmatrix für mehr als zwei Merkmale berechnet werden, so ist diese bei einer Available-Case-Analyse nicht notwendigerweise positiv semidefinit (vgl. Little und Rubin, 2002, S. 55). Dies führt zu Problemen bei multivariaten Verfahren, die eine positive Semidefinitheit beispielsweise zur Invertierung der Kovarianzmatrix voraussetzen. Zur Veranschaulichung, welche der Widersprüche durch eine nicht positiv definite Kovarianzmatrix auftreten können, bedienen sich Little und Rubin (2002, S. 55) folgendes plakativen Beispiels:

Beispiel 2.19: *Widerspruch bei der Analyse verfügbarer Objekte*
Gegeben sei die folgende Datenmatrix, bei der die fehlenden Werte mit ∘ an-

gedeutet sind:

$$A = \begin{pmatrix} 1 & 1 & \circ \\ 2 & 2 & \circ \\ 3 & 3 & \circ \\ 4 & 4 & \circ \\ 1 & \circ & 1 \\ 2 & \circ & 2 \\ 3 & \circ & 3 \\ 4 & \circ & 4 \\ \circ & 1 & 4 \\ \circ & 2 & 3 \\ \circ & 3 & 2 \\ \circ & 4 & 1 \end{pmatrix} .$$

Mittels der direkten Berechnung der Korrelationen unter der Verwendung der Vorgehensweise von Matthai (1951) entsteht folgende Kovarianzmatrix:

$$\mathcal{P} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} .$$

Direkt ist ein Widerspruch in der Kovarianzmatrix zu erkennen. Merkmale 1 und 2 sowie 1 und 3 sind perfekt gleichläufig. Merkmal 2 müsste somit auch gleichläufig mit Merkmal 3 sein. Tatsächlich deutet die mittels Available-Case-Analyse berechnete Korrelation darauf hin, dass die Merkmale 2 und 3 perfekt gegenläufig sind.

Zwar werden bei einer Available-Case-Analyse mehr der vorhandenen Daten genutzt als bei der Complete-Case-Analyse, jedoch führt dies trotzdem nicht immer zu besseren Ergebnissen. In Simulationsstudien mit MCAR fehlenden Daten[17] kommen sowohl Azen und van Guilder (1981, S. 54) als auch Haitovsky (1968, S. 79 f.) zum Entschluss, dass die Available-Case-Analyse nicht immer zu besseren Parameterschätzungen führt. Gemessen an den Problemen, welche durch wechselnde Berechnungsgrundlagen bei der Available-Case-Analyse entstehen (vgl. Beispiele 2.18 und 2.19), erscheint ein Nutzenzuwachs gegenüber der Complete-Case-Analyse nicht gegeben zu sein.

[17] Beide Verfahren sind lediglich beim Vorhandensein dieses Ausfallmechanismus anwendbar.

2.4.2 Imputationsverfahren

Imputation ist die Ersetzung der fehlenden Werte einer Datenmatrix durch Schätzungen. Imputationsverfahren sind somit jene Verfahren, die auf systematische Art und Weise die Werte generieren, welche eine Datenmatrix vervollständigen. Gegenüber den bereits betrachteten Verfahren zeichnen sich Imputationsverfahren insbesondere dadurch aus, dass durch ihre Anwendung eine vollständige Datenmatrix mit denselben Dimensionen wie die ursprüngliche Datenmatrix entsteht. Ähnlich wie bei den Eliminierungsverfahren werden die Imputationsverfahren mit dem Ziel verwendet, statistische Methoden anwenden zu können, die für vollständige Datenmatrizen konzipiert wurden. Entgegen der Eliminierungsverfahren muss jedoch bei den Imputationsverfahren nie auf einen Teil der vorhandenen Informationen verzichtet werden.

Eine Möglichkeit, Imputationsverfahren zu systematisieren, zeigen Little und Rubin (2002, S. 59 ff., vgl. Abbildung 2.18) auf. Demnach entsprechen Imputationswerte entweder einem bestimmten Lageparameter der unterstellten Verteilung von A^{mis} oder einzelnen Werten, die aus der unterstellten Verteilung von A^{mis} gezogen werden (vgl. Little und Rubin, 2002, S. 59). Zudem existieren zwei Möglichkeiten, die Verteilung von A^{mis} aufzustellen. $f\left(A^{mis}\right)$ kann explizit oder implizit modelliert werden. Bei der expliziten Modellierung wird auf ein formales statistisches Modell – wie beispielsweise eine multivariate Normalverteilung – zurückgegriffen. Bei der implizierten Modellierung liegt der Fokus auf dem der Methode zugrundeliegenden Algorithmus.

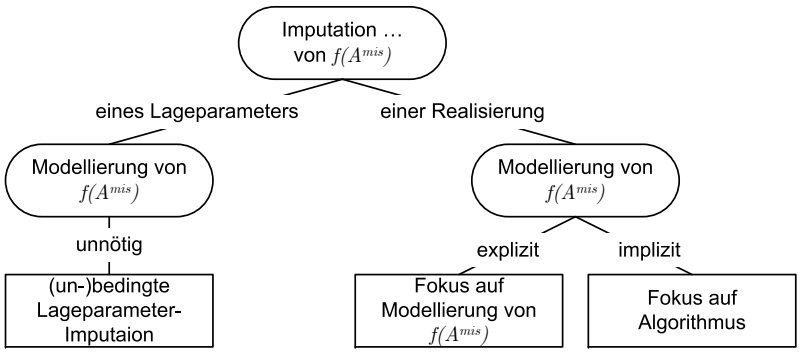


Abbildung 2.18: Die Systematisierung existierender Imputationsverfahren nach Little und Rubin (2002, S. 59)

Die Frage danach, wie die unterstellte Verteilung von A^{mis} zur Erzeugung von Imputationswerten verwendet wird, ist eng mit einer weiteren Eigenschaft

von Imputationsverfahren verwoben, welche zu ihrer Systematisierung genutzt wird. Grundsätzlich können Imputationsverfahren stochastisch oder deterministisch sein. Während deterministische Verfahren sich dadurch kennzeichnen, dass sie bei der Anwendung auf dieselbe Datenmatrix dieselben Imputationen vornehmen, können stochastische Verfahren bei wiederholter Anwendung unterschiedliche Vervollständigungen der Datenmatrix erzeugen. Es ist diese Eigenschaft der stochastischen Verfahren, die eine Sensitivitätsbetrachtung von weiterführenden Analyseergebnissen in Abhängigkeit von den Imputationen erlaubt. Diese Art der Sensitivitätsbetrachtung ist allgemein als „Multiple Imputation“ bekannt (Rubin, 1977; Rubin, 1978). Differenzierungen von Imputationsmethoden, die auf einer Unterscheidung zwischen „Multiple Imputation“ und Methoden, die lediglich einen einzelnen Wert zur Imputation für jeden fehlenden Wert erzeugen (single imputation methods), existieren auch in der Literatur (vgl. etwa Enders, 2010; McKnight et al., 2007). Eine solche Klassifikation kann jedoch nicht als sinnvoll erachtet werden, da jede Imputationsmethode mit stochastischer Komponente lediglich einmal durchgeführt werden kann und somit einer „Single Imputation“ entspricht[18]. „Multiple Imputation“ stellt demnach eher eine Metaebene für die Durchführung von stochastischen Imputationsmethoden dar.

Eine weitere Möglichkeit, Ersetzungsverfahren zu systematisieren, wird von Schnell (1986) angeboten. Schnell (1986) gliedert Imputationsverfahren insgesamt in drei Klassen (vgl. Abbildung 2.19). Zum einen wird zwischen „MD-

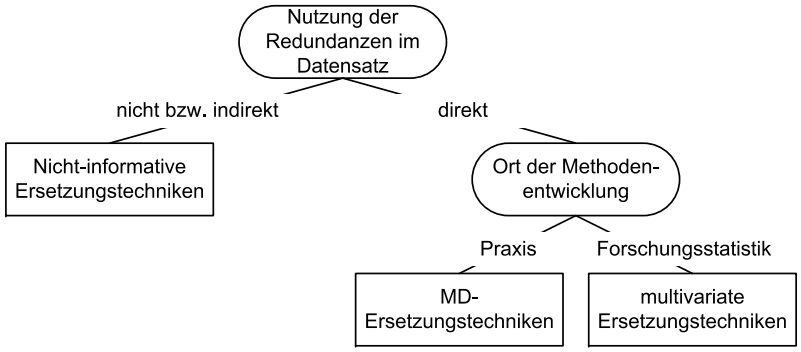


Abbildung 2.19: Die Systematisierung existierender Imputationsverfahren nach Schnell (1986, S. 92, 95, 97, 113)

[18] Deutlich wichtiger ist hier die Unterscheidung zwischen den multiplen Imputationsmethoden, die „proper“ (Rubin, 1987, S. 118) und jenen, die es nicht sind.

Ersetzungstechniken“ (Missing-Data-Ersetzungstechniken) und „multivariaten Ersetzungstechniken“ unterschieden, wobei „MD-Ersetzungstechniken“ in „informative Ersetzungstechniken“ und „nicht-informative Ersetzungstechniken“ unterteilt werden (vgl. Schnell, 1986, S. 92, 95, 97, 113). Gemäß Schnell (1986, S. 95) sind unter den „Nicht-informative Ersetzungstechniken“ Imputationsverfahren zu verstehen, die „. . . nicht auf direkte Weise versuchen, die im Datensatz vorhandene Redundanz zur Ersetzung zu nutzen. . . “. „Informative Ersetzungstechniken“ und „multivariate Ersetzungstechniken“ sind demnach Methoden, die direkt auf jene Redundanzen, welche in einem zu imputierenden Datensatz noch vorhanden sind, zurückgreifen. Die Unterscheidung der Verfahrensgruppen, welche laut Schnell (1986, S. 97) rein willkürlich erfolgt, richtet sich nach der historischen Entwicklung. „Informative Ersetzungstechniken“ stammen vorwiegend aus der Praxis der amtlichen Statistik. Die „multivariaten Ersetzungstechniken“ wurden in der angewandten Forschungsstatistik entwickelt.

Da es sich bei Imputationsverfahren immer um Methoden handelt, welche – mehr oder weniger – mittels vorhandener die fehlenden Informationen ersetzten, ist grundsätzlich jedes Prognose- oder Modellierungsverfahren zur Ermittlung von Schätzwerten möglich. Eine umfassende Klassifikaton müsste somit nicht nur alle Verfahren in Gruppen einteilen, die zur Ermittlung von Imputationswerten verwendet werden, sondern verwendet werden können. Neuere Werke (wie etwa de Waal et al., 2011) beschränken sich daher auf eine Darstellung häufig verwendeter Imputationsmethoden. Dieser Darstellungsform wird sich im Folgenden angeschlossen, wobei jedoch für jede dargestellte Imputationsmethode eine Einordnung in die Systematisierungen nach Schnell (1986, S. 92, 95, 97, 113) beziehungsweise Little und Rubin (2002, S. 59) vorgenommen wird.

2.4.2.1 Imputation eines Lageparameters

Sollen alle fehlenden Werte eines Merkmals durch einen skalenadäquaten Lageparamter ersetzt werden, wird von einer Lageparameterimputation gesprochen. Da in der Missing-Data-Literatur tendenziell der Fokus auf dem Umgang mit kardinalen Merkmalen liegt, wird meist die Mittelwertimputation als Spezialfall der Lageparameterimputation thematisiert. Neben der Verwendung des

arithmetischen Mittelwertes als Schätzwert für die fehlenden Daten sind in Abhängigkeit der Merkmalsskalierung sowohl der Modus, der Median und das geometrische Mittel als auch speziellere Erwartungswerte, wenn eine gewisse Verteilungsannahme getroffen werden kann, denkbar (vgl. Bankhofer, 1995, S. 106).

Die Berechnung des gewählten Lageparameters erfolgt lediglich anhand der vorhandenen Merkmalsausprägungen des Merkmals, welches es zu imputieren gilt; ein Ansatz, welcher auf die Arbeit von Wilks (1932) zurückgeht und garantiert, dass der Mittelwert der imputierten Datenmatrix im Vergleich zu einer Complete-Case-Analyse konstant bleibt. Dies gilt analog für jegliche andere Lageparameter, stellt jedoch lediglich beim Vorliegen eines MCAR-Ausfallmechanismus einen Vorteil dar. Bei MAR- und NMAR-Datenausfall kann nicht davon ausgegangen werden, dass die Lageparameter eines Merkmals ohne Bias direkt mittels der vollständig vorhandenen Ausprägungen geschätzt werden können.

Während logische Inkonsistenzen, die bei der Available-Case-Analyse auftreten können, durch eine Lageparameterimputation ausgeschlossen werden, weist diese dennoch weitere Nachteile auf. Zum einen wird durch die Imputation eines einzelnen Wertes die Variabilität innerhalb eines imputierten Merkmals verringert. Zum anderen reduziert eine Lageparameterimputation die Abhängigkeiten zwischen imputierten Merkmalen. Little und Rubin (2002, S. 61) zeigen beispielsweise für eine Mittelwertimputation, dass selbst unter dem günstigsten MCAR-Ausfallmechanismus die Varianz eines bestimmten Merkmals k um den Faktor $1 - (\tilde{v}_{\bullet k}^{mis} \cdot n - 1)/(n-1)$ unterschätzt wird. Auch die Kovarianz zweier Merkmale wird in Abhängigkeit jener Ausprägungen, die für beide Merkmale vorhanden sind, in ähnlicher Weise unterschätzt.

Beispiel 2.20: *Auswirkungen einer Mittelwertimputation*
Gegeben sei das Merkmal Gesamtfamilieneinkommen aus dem American Community Survey (ACS) (Ruggles et al., 2010; U.S. Bureau of the Census, 2010). Insgesamt liegen Daten zu 2.563.935 Personen vor, bei denen 542.743 keine Angaben zum Gesamtfamilieneinkommen gemacht haben. Dies entspricht circa einem Anteil von 21% fehlenden Werten in diesem Merkmal. In der folgenden Tabelle sind die mittels einer Complete-Case-Analyse und nach einer Mittelwertimputation geschätzten Mittelwerte und Varianzen dargestellt:

	Complete-Case-Analyse	Mittelwertimputation
Mittelwert	35.145,41	35.145,41
Varianz	2.670.065.508	2.104.856.199

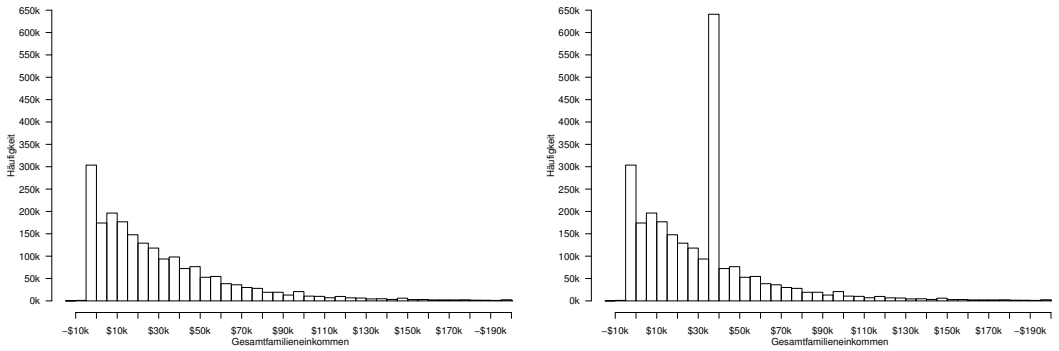
Zu erkennen ist, dass der Mittelwert bei beiden Verfahren gleich und dass die Varianz genau um folgenden Faktor bei der Mittelwertimputation kleiner ist:

$$\frac{2.104.856.199}{2.670.065.508} = 0{,}7883 \; .$$

Dies entspricht auch jenem Faktor, der sich durch Anwendung der Formel von Little und Rubin (2002, S. 61) ergibt:

$$1 - \frac{(0{,}2116 \cdot 2.563.935) - 1}{2.563.935 - 1} = 0{,}7883 \; .$$

Abbildung 2.20 zeigt die offensichtliche Wirkung der Mittelwertimputation auf die mittels eines Histogramms dargestellte eindimensionale Häufigkeitsverteilung. In Abbildung 2.20b ist der Mittelwert beider Verteilungen deutlich zu erkennen. Auch direkt ersichtlich ist, dass die Spitze in der Mitte des Histogramms aus Abbildung 2.20b größer ist als der entsprechende Balken in Abbildung 2.20a. Es sind die 542.743 Objekte, zu denen keine Angaben vorliegen.



(a) Eliminierung (b) Mittelwertimputation

Abbildung 2.20: Vergleich eindimensionaler Häufigkeitsverteilungen nach einer Eliminierung fehlender Werte und einer Mittelwertimputation; Gesamtfamilieneinkommen aus dem American Community Survey (ACS, Ruggles et al., 2010; U.S. Bureau of the Census, 2010), ca. 21% fehlende Werte

Eine Lageparameterimputation muss nicht unbedingt merkmalsweise durchgeführt werden. Insbesondere für die Mittelwertimputation werden in der Literatur zwei weitere Vorgehensweisen vorgeschlagen. Die Möglichkeit, objektweise vorzugehen, wird beispielsweise von Schafer und Graham (2002, S. 157 f.), Enders (2010, S. 50 f.) sowie Graham (2012, S. 51 f.) thematisiert. Hier wird der Mittelwert, welcher zur Ersetzung von fehlenden Werten genutzt wird, über alle vorhandenen Merkmalsausprägungen eines Objekts berechnet. Einen Mittelwert über alle vorhandenen Merkmalsausprägungen der gesamten Datenmatrix zu berechnen und als Imputationswert zu verwenden, wird von Dear (1959) vorgeschlagen. Beide Alternativen sind deutlich weniger erforscht als die Einzelbetrachtung eines jeden Merkmals, wohl, weil gewisse Nachteile dieser Vorgehensweisen offensichtlich sind. Zum einen sollten alle Merkmale, die zur gemeinsamen Mittelwertsberechnung verwendet werden, einen inhaltlichen Bezug zueinander aufweisen und sich die Ausprägungen in einer ähnlichen Größenordnung bewegen. Zum anderen muss eine Datenmatrix ausschließlich Merkmale mit gleicher Skalierung aufweisen. Entsprechend dürfen bei einer Mittelwertimputation nur kardinale Merkmale vorliegen (vgl. Bankhofer, 1995, S. 108).

2.4.2.2 Verhältnisschätzerimputation

Die Verhältnisschätzerimputation entspricht im Grundgedanken einem Dreisatz und erfordert zur Imputation der fehlenden Werte eines kardinalen Merkmals eine Auxiliarvariable. Diese Hilfsvariable l muss zudem mindestens in allen Objekten vollständig sein, bei denen im zu imputierenden Merkmal k Werte fehlen. Zur Bestimmung der Imputationswerte ist zunächst das Verhältnis α_{kl} (ratio) zwischen den Merkmalen k und l wie folgt zu berechnen (vgl. de Waal et al., 2011, S. 244; Shao, 2000, S. 79; Shao und Sitter, 1996, S. 1280):

$$\alpha_{kl} = \frac{\sum_{i=1}^{n-r} a_{ik}^{cc}}{\sum_{i=1}^{n-r} a_{il}^{cc}}. \tag{2.19}$$

Nach der Berechnung dieses Verhältnisses werden die fehlenden Werte in k mittels folgender Formel geschätzt:

$$\hat{a}_{ik} = \alpha_{kl} \cdot a_{il} \; \forall i : v_{ik} = 1 \tag{2.20}$$

Anhand der Formel (2.20) ist ersichtlich, dass zur Schätzung der $\hat{a}_{ik}$ eine lineare Gleichung ohne Absolutglied verwendet wird. Somit handelt es sich bei dieser Multiplikatormethode um eine Art vereinfachte Regressionsimputation (vgl. de Waal et al., 2011, S. 245).

Beispiel 2.21: *Verhältnisschätzerimputation bei der 2009 ACS Stichprobe*
Für die Datenmatrix des Anhangs A (Fall 4) sollen die fehlenden Werte im Merkmal FTOTINC (Gesamtfamilieneinkommen) unter Verwendung eines Verhältnisschätzers imputiert werden. Als Auxiliarvariable wird das Merkmal INCTOT (Individualeinkommen), welches bei einer Complete-Case-Analyse eine Korrelation von 0,749 zu FTOTINC aufweist[19], verwendet. Der Verhältnisschätzer berechnet sich unter Ausschluss der Objekte 4, 5, 7, 8, 13, 15, 20 und 23 bei Verwendung der Formel (2.19) wie folgt:

$$\alpha_{10\,9} = \frac{57.000 + \ldots + 21.000}{13.000 + \ldots + 21.000} = \frac{1.152.724}{590.424} = 1{,}9523.$$

Mittels $\alpha_{10\,9}$ werden dann folgende Imputationswerte berechnet:

$$\begin{aligned} \hat{a}_{4\,10} &= \alpha_{10\,9} \cdot a_{4\,9} &&= 1{,}9523 \cdot 40.000 = 78.094{,}66 \\ &\vdots && \vdots \\ \hat{a}_{23\,10} &= \alpha_{10\,9} \cdot a_{23\,9} &&= 1{,}9523 \cdot 35.000 = 68.332{,}83. \end{aligned}$$

Eine Imputation mittels eines Verhältnisschätzers wird zumeist als einfaches Wachstumsmodell verwendet. Hierbei entspricht dann das Merkmal k dem Merkmal l zu einem späteren Zeitpunkt (vgl. de Waal et al., 2011, S. 244). Somit wäre die hohe Korrelation zwischen k und l aufgrund des zeitlichen Zusammenhangs gesichert.

Denkbar ist auch, dass das Verhältnis α_{kl} mittels einer Available-Case-Analyse berechnet wird. In diesem Fall müssten der Nenner und der Zähler um Eins durch die Anzahl der jeweiligen Summanden korrigiert werden. Hierdurch würde das Verhältnis der Mittelwerte genutzt werden.

[19] Die tatsächliche Korrelation beträgt 0,565.

Genauso einfach wie dieses Verfahren in der Durchführung ist, genauso offensichtlich sind seine Schwächen. Lässt sich der Zusammenhang zwischen k und l nicht rein durch einen Verhältnisunterschied erklären, so ist diese Methode ungeeignet. Insbesondere wenn der Zusammenhang zwischen den Merkmalen nicht linear ist, kann die Verwendung des Verhältnisschätzers zu schlechter Imputationsqualität führen. Zwar ist es denkbar, dass diese Methode beim Vorliegen eines MAR-Ausfallmechanismus anwendbar ist, dennoch ist es unklar, wie hoch die Korrelation zwischen k und seinem Hilfsmerkmal sein muss, um akzeptable Imputationswerte zu erzielen.

2.4.2.3 Regressionsimputation

Bei einer Regressionsimputation werden die fehlenden Werte in einem Merkmal mit Hilfe der Schätzungen aus einem Regressionsmodell ersetzt. Diese auch als bedingte Mittelwertimputation bekannte Methode (vgl. Little und Rubin, 2002, S. 62 ff.; Enders, 2010, S. 44) verwendet die Informationen aus einer oder mehr Kovariaten[20]. In seinen Annahmen stellt diese Methode die natürliche Erweiterung der Mittelwertimputation dar.

Die einfachste Form der Regressionsimputation betrachtet jedes Merkmal mit fehlenden Daten sequentiell und schätzt die nötigen Regressionskoeffizienten durch die Methode der kleinsten Quadrate (MKQ) mittels einer Complete-Case-Analyse (vgl. Little und Rubin, 2002, S. 62; Enders, 2010, S. 44; de Waal et al., 2011, S. 237). Alternativ kann auch objektweise wie in der Methode von Buck (1960) vorgegangen werden. Hier werden für jedes der vorhandenen Ausfallmuster so viele Regressionsgleichungen geschätzt wie fehlende Werte bei dem entsprechenden Ausfallmuster vorhanden sind. Für jede Schätzung wird auf jene Objekte zurückgegriffen, die in allen Merkmalen zusätzlich zu jenem Merkmal, bei dem imputiert werden soll, vollständig sind.

Beispiel 2.22: *Regressionsimputation bei der 2009 ACS Stichprobe*
Für die Datenmatrix des Anhangs A (Fall 3 und 4) sollen die fehlenden Werte in den Merkmalen INCTOT (a_{-9}) und FTOTINC (a_{-10}) mittels Regression im-

[20] Im Falle dessen, dass keine Kovariaten verwendet werden, entspricht die Regressionsimputation einer einfachen Mittelwertimputation (unconditional mean imputation, vgl. Little und Rubin, 2002, S. 61; Enders, 2010, S. 42).

putiert werden. Dabei werden die Ergebnisse der sequentiellen Vorgehensweise mit der Vorgehensweise von Buck (1960) verglichen. Als zusätzliche Kovariaten werden die Merkmale AGE ($a_{\_4}$) und UHRSWORK ($a_{\_8}$) eingesetzt. In der sequentiellen Vorgehensweise entstehen zwei verschiedene Regressionsgleichungen unter der Verwendung derselben Objektmenge (Objekte 2, 3, 6, 9, 14, 16, 19, 21, 22, 24 und 25). Im Einzelnen entstehen folgende Regressionsfunktionen:

$$\hat{a}_{i\,9} = -91.042 + 2.389 \cdot a_{i\,4} + 1.151 \cdot a_{i\,8}$$
$$\hat{a}_{i\,10} = -51.132 + 1.659 \cdot a_{i\,4} + 2.439 \cdot a_{i\,8}.$$

Für die Vorgehensweise nach Buck (1960) werden vier Gleichungen aufgestellt. Die obigen Gleichungen finden Anwendung für jene Objekte, bei denen Werte in den Merkmalen INCTOT und FTOTINC fehlen. Für die beiden Fälle, dass lediglich eine Merkmalsausprägung bei INCTOT beziehungsweise FTOTINC fehlt, werden die folgenden Regressionsfunktionen geschätzt:

$$\hat{a}_{i\,9} = -60.812 + 1.408 \cdot a_{i\,4} - 290 \cdot a_{i\,8} + 0{,}5912 \cdot a_{i\,10}\ \forall i : v_{i\,9} = 1 \wedge v_{i\,10} = 0$$
$$\hat{a}_{i\,10} = 35.580 - 616 \cdot a_{i\,4} + 1.342 \cdot a_{i\,8} + 0{,}9524 \cdot a_{i\,9}\ \forall i : v_{i\,9} = 0 \wedge v_{i\,10} = 1.$$

Durch diese Ausführungen wird klar, dass die Methode nach Buck (1960) mehr Informationen verwendet, sofern diese vorhanden sind.

Grundsätzlich ist es das Ziel einer multiplen Regression, mit Hilfe mehrerer unabhängiger, kardinal skalierter Merkmale ein abhängiges Merkmal mit gleicher Skalierung zu erklären. Diese Annahmen können jedoch an sich im Einzelnen relaxiert werden. So können auch nominale Merkmale als unabhängige Merkmale dienen, wenn sie geeignet umkodiert werden[21]. Werden lediglich umkodierte nominale Merkmale zur Erklärung eines kardinal skalierten Merkmals verwendet, entspricht die Regressionsimputation einer Imputation mittels Varianzanalyse (vgl. Bankhofer, 1995, S. 134 ff.). Auch die abhängige Variable, das Merkmal, welches es zu imputieren gilt, kann nominal skaliert sein. Bei einem dichotomen Merkmal entspricht dies bei der Anwendung des MKQ-Ansatzes einem linearen Wahrscheinlichkeitsmodell (vgl. Jobson, 1992, S. 282 f.). Alternativ lassen sich hier auch Logit- beziehungsweise Probit-Modelle aufstellen. In diesen Fällen werden die Regressionskoeffizienten, wegen der nicht-

[21] Kodierungsverfahren, die im Falle einer Regression grundsätzlich geeignet sind, werden beispielsweise von Fahrmeir et al. (1996b, S. 94) dargestellt.

linearen Linkfunktion, nicht unter Nutzung des MKQ-Ansatzes geschätzt, sondern es wird ein iterativer Algorithmus verwendet (vgl. Jobson, 1992, S. 285 ff.). Soll ein nominal-polytomes Merkmal imputiert werden, kann beispielsweise eine multinomiale logistische Regression (vgl. Jobson, 1992, S. 306 ff.) zur Identifikation der Imputationswerte verwendet werden. In diesen Fällen kann von einer bedingten Modusimputation gesprochen werden. Ist die abhängige Variable ordinal oder kardinal skaliert, lässt sich zudem eine bedingte Medianimputation mittels einer Quantilregression (Koenker und Bassett, 1978) realisieren.

Unabhängig davon, welche Skalierung die zu imputierenden und die zur Hilfe verwendeten Merkmale aufweisen, existieren neben dem sequentiellen Vorgehen und dem Vorgehen nach Buck (1960) weitere Verfahrensvarianten. Diese unterscheiden sich im Wesentlichen von den Grundvarianten dadurch, dass sie mehr der vorhandenen Ausprägungen nutzen oder bereits imputierte Werte mit in die Schätzungen einfließen lassen. Einen umfassenden Überblick weiterer Verfahrensvarianten und wenig verbreiteter Ansätze zeigt Bankhofer (1995, S. 126 ff.) auf.

2.4.2.4 Deck-Verfahren

Soll die Merkmalsausprägung, welche in einem anderen Objekt vorhanden ist, als Imputationswert dienen, wird entweder ein Hot-Deck-Verfahren oder ein Cold-Deck-Verfahren angewandt. Hierbei sind unter Hot-Deck-Verfahren jene Imputationsmethoden zu subsumieren, bei denen die fehlenden Werte in einem Objekt durch die Verdopplung der vorhandenen Werte eines anderen, ähnlichen Objekts aus derselben Datenmatrix ersetzt werden, so dass ein Datensatz ohne Datenausfall erzeugt wird (vgl. Ford, 1983, S. 186; Sande, 1983, S. 341). Bei Cold-Deck-Verfahren werden hingegen Objekte zur Verdoppelung verwendet, die aus einer anderen Datenmatrix stammen (vgl. Ford, 1983, S. 186; Shao, 2000, S. 80). Es erfolgt also ein Zugriff auf eine der Analyse externe Datenquelle zur Generierung geeigneter Imputationswerte (Little und Rubin, 2002, S. 60). Beide Alternativen basieren jedoch auf den impliziten Modellannahmen, dass die Verteilung jener als ähnlich identifizierten Objekte der Verteilung von A^{mis} entspricht (vgl. Little und Rubin, 2002, S. 66). Somit ist auch die Effektivität dieser Verfahren an die Erfüllung von zwei Sachverhalten

gebunden. Zum einen muss in der Datenquelle, die zur Suche nach einem Objekt mit geeigneten Merkmalsausprägungen verwendet wird, auch ein solches Objekt existieren. Zum anderen muss das Objekt, welches diese geeigneten Merkmalsausprägungen aufweist, mittels der konkreten Methode identifiziert werden können.

Grundsätzlich sind, bis auf die Quelle der Objekte, welche zur Verdoppelung der Werte herangezogen wird, identische Vorgehensweisen zur Identifikation eines geeigneten Objekts denkbar. Rein faktisch wird jedoch die Cold-Deck-Imputation lediglich angewendet, wenn eine frühere Untersuchung zu denselben Sachverhalt als externe Datenquelle zur Verfügung steht (vgl. Schnell, 1986, S. 108). Bei Hot-Deck-Verfahren liegt hingegen ein stärkerer Fokus auf dem Algorithmus; daher werden in der Literatur entsprechend mehr Verfahrensvariationen vorgestellt. Diese unterscheiden sich anhand folgender Fragen:

- Wie wird die Ähnlichkeit zwischen jenen Objekten, die einer Imputation bedürfen und die zu einer Imputation herangezogen werden können, quantifiziert?
- Erfolgt eine deterministische Zuordnung zwischen jenen Objekten, die einer Imputation bedürfen und die zu einer Imputation herangezogen werden können, oder wird ein Verfahren mit Zufallskomponente verwendet?
- Wie wird verfahren, wenn mehrere Merkmale einen Datenausfall aufweisen?
- Wird begrenzt, wie häufig ein einzelnes Objekt zur Verdoppelung seiner Werte dienen kann?

Eine ausführliche Systematisierung und Darstellung der existierenden Hot-Deck-Verfahren erfolgt in Kapitel 3.

Von besonderer Bedeutung wird die Cold-Deck-Imputation, wenn der externe Datensatz Informationen zu einem früheren Zeitpunkt über dieselben Untersuchungsobjekte beinhaltet. In diesem Fall geschieht die Verdoppelung eines früheren Wertes, welches auch unter dem Namen „Last Observation Carried Forward“ bekannt ist (vgl. Enders, 2010, S. 51 f.). Die Verdoppelung der Werte geschieht dann unter der Annahme, dass keine Veränderung zwischen

den Erhebungszeitpunkten stattgefunden hat. Laut de Waal et al. (2011, S. 245) kann dieses Vorgehen auch als Spezialfall der Imputation mittels Verhältnisschätzer betrachtet werden, bei dem das Verhältnis $\alpha_{kl} = 1$ gesetzt wird. Liegt jedoch ein zeitlicher Zusammenhang zwischen den Objekten der zwei Datensätzen vor, erscheint es fragwürdig, ob einer der Datensätze tatsächlich extern zu der Untersuchung ist. Denkbar scheint es auch, dass beide Datensätze als Teil einer Longitudinalstudie zu betrachten sind und es sich somit bei dem „Last Observation Carried Forward“ Verfahren um eine Hot-Deck-Methode handelt. Letztlich ist hierbei darauf abzustellen, ob Daten zu dem vorherigen Zeitpunkt nach der Imputation für weitergehende Analysen verwendet werden. Dienten die Informationen zu den vorherigen Zeitpunkten nur zur Imputation, so handelt es sich um ein Cold-Deck-Verfahren. Werden die Datensätze in weitergehenden Analysen gemeinsam analysiert, so handelt es sich um ein Hot-Deck-Verfahren.

Beispiel 2.23: *Last Observation Carried Forward bei Longitudinaldaten*
Gegeben seien jene in der folgenden Tabelle dargestellten Einkommen von fünf Personen in den Jahren 2011 bis 2014, wobei die ∘ fehlende Werte andeuten. Diese Merkmale stellen jeweils einen Auszug unterschiedlicher Datensätze mit weiteren Merkmalen dar.

	Einkommen in den Jahren			
	2011	2012	2013	2014
Person 1	50.000	53.000	∘	∘
Person 2	47.000	46.000	49.000	51.000
Person 3	43.000	∘	∘	∘
Person 4	55.000	∘	56.000	59.000
Person 5	45.000	45.000	47.000	46.000

Wird nun die letzte Beobachtung fortgeschrieben, so entsteht folgende imputierte Datenbasis:

	Einkommen in den Jahren			
	2011	2012	2013	2014
Person 1	50.000	53.000	53.000	53.000
Person 2	47.000	46.000	49.000	51.000
Person 3	43.000	43.000	43.000	43.000
Person 4	55.000	55.000	56.000	59.000
Person 5	45.000	45.000	47.000	46.000

Durch die mit dem Auswahlprozess assoziierte Duplizierung von Werten, die bereits vorkommen, wird garantiert, dass keine Transformation oder Rundung von Imputationswerten stattfinden muss, um zulässige Imputationswerte zu erzeugen. Des Weiteren bleiben Eigenschaften der Verteilung, wie etwa Unstetigkeiten, erhalten. Im Allgemeinen erfreuen sich die Deck-Verfahren aufgrund ihrer Vielseitigkeit und geringen Implementierungskosten einer hohen Beliebtheit (vgl. Kalton und Kasprzyk, 1982, S. 28). Diesen Vorteilen steht jedoch gegenüber, dass die theoretischen Eigenschaften der Verfahren nicht oder nur unzureichend ergründet sind (vgl. Little und Rubin, 2002, S. 60 f.; Andridge und Little, 2010, S. 40).

Kapitel 3

Hot-Deck-Verfahren

Der Begriff Hot-Deck (heißer Stapel) beschreibt ursprünglich eine Verfahrensweise für den Umgang mit fehlerhaften Lochkarten (vgl. Ono und Miller, 1969, S. 277). Während der Zeit, als die sogenannten Hollerith-Lochkarten im Einsatz waren, wurde auf methodisch simple Verfahren zurückgegriffen, um mit fehlerhaften Lochkarten, die zur Speicherung von Daten dienten, umzugehen. Fehlerhafte Lochkarten wurden durch fehlerfreie ersetzt. Wurde eine fehlerfreie Karte aus dem aktuell auszuwertenden Stapel als Ersatz genutzt, wurde der heiße Stapel verwendet (vgl. Rizvi, 1983, S. 351). Kam hingegen eine Karte aus einer anderen Quelle zum Einsatz, wie etwa einem älteren Datensatz, oder wurde ad hoc eine neue Karte gestanzt, war von einer Cold-Deck-Prozedur die Rede (kalter Stapel, vgl. Little, 1982, S. 244; Rizvi, 1983, S. 351). Um sicherzustellen, dass eine möglichst ähnliche Karte die defekte ersetzt, wurde keine beliebige Karte aus dem heißen Stapel gewählt, sondern die letzte Karte, die fehlerfrei eingelesen werden konnte. Zur weiteren Verbesserung der Ergebnisse wurde der Kartenstapel vor dem Einlesen in den Rechner nach vollständig vorhandenen Eigenschaften, wie etwa der Adresse der befragten Haushalte, sortiert (Huckett und Larsen, 2007, S. 3055).

Da Lochkarten seit dem 1890-Zensus in dem U.S. Bureau of the Census im Einsatz waren, wurden diese auch der Defacto-Standard zur Dateneingabe für alle zukünftigen Erhebungen des Census Bureau. So sollte es dann auch sein, als 1942 die Verantwortung für das Sample Survey of Unemployment, welches seit dem als Current Population Survey (CPS) bekannt ist, zum Census Bureau transferiert wurde. Diese erste Arbeitslosenstatistik der USA, welche

auf Grund der großen Depression und den hiermit verbundenen Unruhen[1] in den 1930er Jahren entwickelt wurde, sollte eine direkte monatliche Messung der Arbeitslosigkeit erlauben, so dass derartige Probleme in der Zukunft nicht wieder auftreten. Mit dem Transfer zu dem U.S. Bureau of the Census kamen auch Änderungen in der Methodologie, welche im Oktober 1943 in Kraft traten (vgl. U.S. Bureau of the Census, 2006, S. 2–1 f.). Teil dieser Änderungen war die Anwendung des im vorherigen Absatz beschriebenen Urtyps des Hot-Deck, welches sicherstellte, dass zumindest die Stichprobengröße pro Frage konstant bleibt. 1960 konnte dann erstmals das klassische sequentielle Hot-Deck zur Imputation fehlender und fehlerhafter Werte auf Grund der nun vorhandenen elektronischen Rechentechnik verwendet werden (Cresce Jr. et al., 2005, S. 2928). 1962 wurde dieses dann auch erstmals beim Zusatzbogen für das Einkommen (income supplement) im CPS verwendet (Oh und Scheuren, 1980, S. 408).

Seither finden Hot-Deck-Verfahren insbesondere im anglo-amerikanischen Raum häufig Anwendung bei der Erstellung von offiziellen Statistiken. So werden verschiedene Hot-Deck-Varianten zur Behebung von fehlenden Werten vom U.S. Census Bureau nicht nur beim CPS eingesetzt, sondern beispielsweise auch beim Survey of Income and Program Participation (SIPP) und American Community Survey (ACS). Auch bei dem britischen Office for National Statistics wurden Hot-Decks zur Imputation im 2001- und 2011-Zensus verwendet. In Kanada werden Hot-Deck-Verfahren bei Statistics Canada sogar in 45% aller aktiven Statistiken, bei denen fehlende Werte auftreten, wie etwa dem Survey of Labour and Income Dynamics (SLID) und der Labour Force Survey (LFS), verwendet (vgl. Joenssen und Bankhofer, 2012, S. 60).

Bereits 1983 kommentierte Ford darauf, dass die Entwicklung von Hot-Deck-Verfahren in der Literatur aufgrund ihres Ursprungs als Praktikerverfahren einige Defizite aufweist. Zu diesen Defiziten zählte damals, dass es keine einheitliche Definition von Hot-Deck-Verfahren gibt (Ford, 1983, S. 185),

[1] Die Arbeitslosigkeit war in den 1930er Jahren in den USA ein derartig großes Problem, dass es zu mehreren Unruhen und einigen pseudo-militärischen Auseinandersetzungen kam. So demonstrierten im Frühling und Sommer 1932 43.000 Arbeitslose, größtenteils Veteranen, und deren Angehörige vor dem Amerikanischen Kongress. Diese als „Bonus Army“ bezeichnete Versammlung wurde durch Kavallerie, Panzer und Infanterie mit Maschinengewehren am 28. Juli 1932 zerschlagen (vgl. Dickson und Allen, 2006, S. 203 f.).

ein Missstand, der bis dato nicht behoben wurde. Durch frühe Entwicklungen im Bereich der Hot-Deck-Imputation entstanden in den hierauf aufbauenden Literatursträngen unterschiedliche Verständnisse des Begriffs „Hot-Deck“. Die Existenz dieser in der Literatur diskutierten unterschiedlichen Definitionen erfordert, dass im folgenden Abschnitt 3.1 zunächst eine Darstellung der unterschiedlichen Definitionen und danach eine Festlegung auf eine Definition stattfindet. Nach der Definition von Hot-Deck-Verfahren erfolgt im Abschnitt 3.2 eine systematische Darstellung der Hot-Deck-Methoden basierend auf ihren Eigenschaften. Das Kapitel schließt mit einer Darstellung von Hot-Deck-Methoden als ganzzahliges Optimierungsproblem ab.

3.1 Definition und Abgrenzung

Grundsätzlich existieren in der Literatur zwei unterschiedliche Verständnisse von dem, was ein Imputationsverfahren zu einem Hot-Deck-Verfahren macht. In der einen Gruppe an Veröffentlichungen wird ein eher allgemeines Verständnis von dem, was ein Hot-Deck-Verfahren ausmacht, vertreten, während in der anderen Gruppe der Begriff „Hot-Deck-Verfahren“ eher restriktiv definiert wird.

Die weiter gefasste und verbreitete Definition von Hot-Deck-Verfahren geht zurück auf Ford (1983) und Sande (1983). Ford (1983, S. 186) legt folgendes Verständnis von Hot-Deck-Verfahren vor:

> „In general, a hot-deck procedure is a duplication process – when a value is missing from a sample, a reported value is duplicated to represent this missing value. [...] The adjective “hot” refers to imputing with values from the current sample.“

Auch die Definition von Sande (1983, S. 341) enthält dieselben wesentlichen Komponenten:

> „We define a hot-deck imputation procedure to be one where an incomplete response is completed by using values from one or more other records on the same file (i.e., from the same survey), and the choice of these records varies with the record requiring imputation.“

Diesem allgemeineren Verständnis schließen sich beispielsweise Schnell (1986, S. 109 f.), Huisman (2000, S. 335), Little und Rubin (2002, S. 66 f.), Dillman et al. (2002, S. 21), Kim und Fuller (2004, S. 559–560), Dahl (2007, S. 5914), Herzog et al. (2007, S. 66–68), Enders (2010, S. 49), Schlomer et al. (2010, S. 4) und de Waal et al. (2011, S. 140–141) explizit an. Weitere Autoren legen dieses Verständnis von Hot-Deck-Verfahren eher implizit zugrunde. So weisen beispielsweise Rao und Shao (1992, S. 811), Chen et al. (2000, S. 1160–1161) oder Chen und Shao (2001, S. 260) darauf hin, dass es sich bei der von ihnen beschriebenen Methode um eine spezielle Form von Hot-Deck-Imputationsmethoden handelt. Des Weiteren findet sich diese Definition in den Reviews von Kalton und Kasprzyk (1986, S. 7) und Andridge und Little (2010, S. 40–41) wieder.

Dem Verständnis dieser Autoren sind drei Komponenten gemein. Erstens sollen zur Behebung fehlender Werte vorhandene Werte verwendet werden. Zweitens sollen die vorhandenen Beobachtungen von Objekten stammen, die in derselben Datenmatrix zu finden sind wie auch die Objekte, bei denen Werte fehlen. Drittens wird betont, dass die vorhandenen Beobachtungen für die Imputation verdoppelt werden, also nicht aggregiert zur Imputation verwendet werden. Demzufolge handelt es sich bei Hot-Deck-Verfahren für diese Autoren um Zuordnungsmethoden. Für diese Zuordnung werden die Objekte in der vorhandenen Datenmatrix in zwei Gruppen aufgeteilt. Jene Objekte, die fehlende Werte aufweisen, daher einer Imputation bedürfen, werden der Gruppe der Empfänger (recipients[2]) und jene Objekte, die Werte spenden und somit zur Imputation herangezogen werden können, werden der Spendergruppe zugeteilt (donors). Diese Aufteilung kann explizit oder implizit durch den Hot-Deck-Algorithmus erfolgen, wobei je nach Algorithmus und vorhandenem Ausfallmuster die Empfängermenge R und Spendermenge D nicht notwendigerweise disjunkt sein müssen. Nach der Zuteilung der Objekte in diese Gruppen werden Spender den Empfängern (gemäß der Vorgaben des konkreten Hot-Deck-Algorithmus) zugeordnet. Die Behebung des Ausfalls beim Empfänger wird durch eine Verdopplung der Werte des Spenders über die fehlenden Werte des

[2] In der Benennung der Empfänger ist die Literatur nicht einheitlich. Vorwiegend wird der Begriff „recipient“ verwendet, jedoch sind auch die Begriffe „donee“ (wie etwa in Siddique und Belin, 2008, S. 81) und „beggar“ (wie etwa in Marker et al., 2002, S. 329) zu finden.

Empfängers erreicht. Weitere prozedurale Details geben die Verfahrensvarianten vor. Dieses Grundprinzip ist in aggregierter Form der Abbildung 3.1 zu entnehmen.

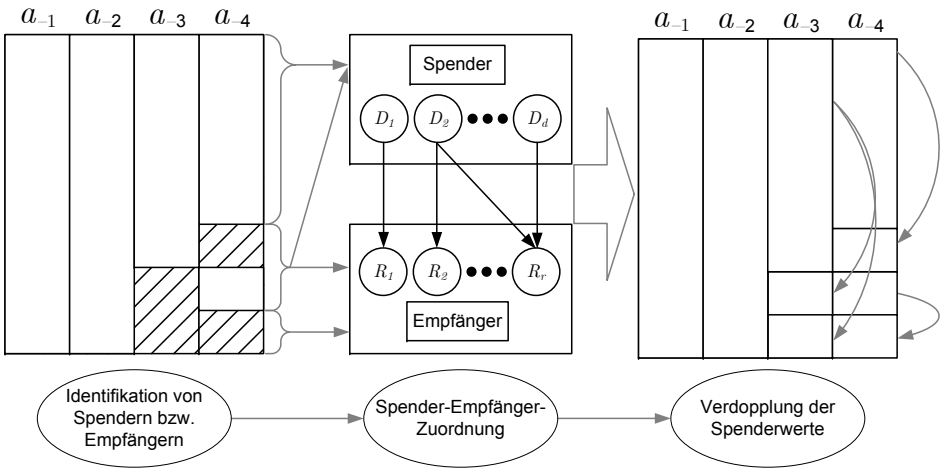


Abbildung 3.1: Grundprinzip der Hot-Deck-Verfahren gemäß dem weiter gefassten Verständnis

Jene Autoren mit restriktiver Definition schränken das obige Verständnis von Hot-Deck-Verfahren ein. So verstehen diese Autoren unter Hot-Deck-Verfahren ein spezielles Verfahren oder eine spezielle Verfahrensvariante. Beispielsweise bezeichnet Fay (1999, S. 213) stets das ursprüngliche sequentielle Verfahren, in dem die Ähnlichkeit zwischen Spender und Empfänger nur durch die Reihenfolge im Datensatz bestimmt wird, als Hot-Deck-Verfahren. Marker et al. (2002, S. 329–330) und Allison (2001, S. 57–58) benennen lediglich Verfahren, bei denen eine zufällige Zuordnung von Spender zu Empfänger innerhalb von Klassen vorgenommen wird, als Hot-Deck-Imputation. McKnight et al. (2007, S. 182–183) sehen solche Imputationsverfahren als Hot-Deck-Verfahren an, die eine zufällige Zuordnung zwischen Spender und Empfänger vornehmen, ob mit oder ohne Verwendung von Klassen. Bankhofer (1995, S. 120 f.) sowie Thran und Gillis (1992, S. 211) stellen fest, dass bei Hot-Deck-Verfahren der ursprüngliche Datensatz vor der Spender-Empfänger-Zuordnung zwangsweise in Klassen aufgeteilt werden muss. Von Durrant und Skinner (2006, S. 6) werden auch Hot-Deck-Verfahren als zufällige Auswahlverfahren dargestellt. Die Abgrenzung zu den anderen Zuordnungsverfahren, die auf einer Distanzfunktion basieren, geschieht eher implizit. Auch Roth (1994, S. 543–544) schließt

sich in einem Review-Artikel durch eine Vorgehensbeschreibung implizit dieser eingeschränkten Sichtweise an.

Bei konkurrierenden Verständnissen dieser Art ist es erforderlich, eine Auswahl zu treffen und diese für den weiteren Verlauf der Arbeit zugrunde zu legen. Grundsätzlich besteht die Wahlmöglichkeit zwischen der eher allgemein gehaltenen Definition und einer der spezielleren Verständnisse. Autoren, die Hot-Deck-Verfahren als ein eher spezielles Verfahren sehen, bilden keine einheitliche Linie. So ist es eher selten, dass zwei Autoren exakt dieselbe Auffassung von dem, was ein Hot-Deck ist, haben. Des Weiteren sind manche Autoren in diesem Bereich nicht einheitlich mit der Definition innerhalb ihrer eigenen Werke. Beispielsweise sieht Longford (2005, S. 43–44) Imputationsverfahren, die eine Distanzfunktion verwenden, zunächst nicht als Hot-Deck-Verfahren an. Später bezeichnet Longford (2005, S. 75) diese Nearest-Neighbor-Verfahren doch als einen Spezialfall der Hot-Deck-Verfahren. Ferner ist auch die Terminologie eher bei jenen Autoren, die der weiteren Definition folgen, einheitlich. Die Unschärfe in diesem Bereich ist nicht neu und wurde neben Ford (1983, S. 185) bereits auch von Schnell (1986, S. 109) bemängelt. Dieser merkte an:

> „Die Schwierigkeit einer Definition der Hot-Deck-Techniken resultiert aus den fließenden Übergängen der Verfahren: Fast jede Technik besitzt als „degenerierten“ Spezialfall zumindest eine Methode, die üblicherweise nicht als Hot-Deck-Technik bezeichnet wird; [. . .]“

Diese spiegelt sich auch in neuerer Literatur wider. So ist auch zu beobachten, dass die distanzbasierten Zuordnungen nicht einheitlich unter Nearest-Neighbor-Verfahren geführt werden. Sie werden gelegentlich als „Donor Imputation“ bezeichnet (vgl. Beaumont und Bissonnette, 2011, S. 173). Ebenfalls wird das ursprüngliche sequentielle Verfahren, welches rein auf der Objektsortierung basiert, als Nearest-Neighbor-Hot-Deck bezeichnet (vgl. West et al., 1990, S. 255). Auch ist in den letzten Jahren in den Veröffentlichungen eher eine Konvergenz auf die allgemeinere Definition zu beobachten.

Demnach wird die nachstehende Definition von Hot-Deck-Imputationsverfahren für die verbleibende Arbeit unterstellt:

Hot-Deck-Verfahren sind Imputationsmethoden, bei denen die fehlenden Werte in einem Objekt durch die Verdopplung der vorhandenen Werte eines anderen, ähnlichen Objekts aus derselben Datenmatrix ersetzt werden, so dass ein Datensatz ohne Datenausfall erzeugt wird.

3.2 Varianten der Hot-Deck-Methoden

Anhand der in dieser Arbeit zugrunde gelegten Definition können die Annahmen, unter denen Hot-Deck-Verfahren allgemein valide Imputationsergebnisse erzeugen, hergeleitet werden. Hot-Deck-Methoden setzen voraus, dass innerhalb der vorliegenden Datenmatrix für jeden Empfänger mindestens ein Spender existiert, dessen Merkmalsausprägungen den unbeobachteten Werten des Empfängers hinreichend nahe kommen. Zudem muss angenommen werden, dass diese Spender mittels der gewählten Methode identifiziert werden können. Grundsätzlich variieren die in der Literatur beschriebenen Hot-Deck-Methoden in den Randbedingungen darin, wie passende Spender identifiziert werden. Diese Varianten lassen sich anhand der folgenden vier zentralen Eigenschaften unterscheiden:

1. Definition von Ähnlichkeit
2. Stochastizität
3. Behandlung mehrerer Merkmale
4. Mehrfachverwendung der Spender

Diese Eigenschaften sowie deren existente Ausprägungen werden in den folgenden vier Unterabschnitten näher beschrieben.

3.2.1 Definition von Ähnlichkeit

Die für die Identifizierung eines ähnlichen Spenders wichtigste Komponente eines Hot-Deck-Verfahrens stellt die Wahl des Ähnlichkeitsmaßes dar. Zwar lässt grundsätzlich die in Abschnitt 3.1 aufgeführte Definition zu, dass einem Empfänger ein beliebiger Spender zugewiesen wird; dieses führt jedoch nur

zu sinnvollen Imputationen, sofern unterstellt werden kann, dass ein beliebiges anderes Objekt innerhalb der Datenmatrix dem Empfänger ähnlich ist. Laut Andridge und Little (2010, S. 49) ist das lediglich beim Vorliegen eines MCAR-Ausfallmechanismus der Fall (vgl. Abschnitt 2.3.1). Wenn die MCAR-Annahme nicht zutrifft oder tiefergreifende Analysen, bei denen die konkreten Werte eines Objekts von Interesse sind, durchgeführt werden sollen, führt die Auswahl eines beliebigen Spenders zu verzerrten Imputationsergebnissen. Grundsätzlich sollte deshalb, weil die MCAR-Annahme meist unrealistisch ist (Little, 1992, S. 1229) und die Ersetzung der fehlenden Werte eines Objekts durch die Ausprägungen eines ähnlichen grundsätzlich sinnvoller erscheint (vgl. Bankhofer, 1995, S. 120), kein beliebiger Spender gewählt werden.

Zur Identifikation der für einen Empfänger hinreichend ähnlichen Spender existieren bei Hot-Deck-Verfahren grundsätzlich drei Varianten. In der Literatur werden Klassenbildungsverfahren, Distanzmaße sowie Objektsortierungsverfahren diskutiert und verwendet. Diese und die Möglichkeit, diese zu kombinieren, werden nun näher erläutert.

3.2.1.1 Klassenbildung

Eine Möglichkeit ähnliche Spender für die Empfänger zu bestimmen, ist eine Aufteilung der gesamten Daten in möglichst homogene Gruppen (vgl. Bankhofer, 1995, S. 121). Diese Homogenität innerhalb der gebildeten Klassen soll garantieren, dass die Spender und Empfänger der gleichen Verteilung folgen (vgl. Ford, 1983, S. 186). Eine solche Zielsetzung erfordert jedoch, dass die zur Erstellung der Klassen verwendeten Variablen zwei Bedingungen erfüllen. Zum einen sollten die verwendeten Merkmale einen starken Zusammenhang zu den beobachteten Werten der zu imputierenden Merkmale aufweisen. Zum anderen sollten Merkmale, mittels derer die Klassifikation erstellt wurde, auch einen starken Zusammenhang zu den nicht beobachteten Werten der zu imputierenden Merkmale besitzen (vgl. Ford, 1983, S. 186). Während die erste Bedingung mittels der vorhandenen Daten überprüfbar ist, muss der zweiten Forderung mit theoretischen Überlegungen und Expertenwissen begegnet werden, so dass keine irreführenden Ergebnisse durch die Imputation erzeugt werden (vgl. Ford, 1983, S. 186).

Grundsätzlich ist jedes existierende Klassifikationsverfahren zur Aufteilung

der Spender und Empfänger in Imputationsklassen anwendbar. Da aber eine Vielzahl an existierenden Klassifikationsalgorithmen im Kontext der Hot-Deck-Verfahren keine Erwähnung finden, wird sich im folgenden Abschnitt auf jene Verfahren beschränkt, die für Hot-Deck-Verfahren eine übergeordnete Bedeutung haben. Eine ausführliche Übersicht an Klassifikationsverfahren bieten beispielsweise Jobson (1992) und Fahrmeir et al. (1996a).

Adjustment-Cell-Methode

Die in der Praxis wohl beliebteste Methode zur Erzeugung von Imputationsklassen ist die Adjustment-Cell-Methode (vgl. Little, 1988, S. 289 oder Andridge und Little, 2010, S. 42). Bei der Adjustment-Cell-Methode werden Imputationklassen anhand der Kreuzklassifikation aller Merkmalsausprägungen der vollständig vorhandenen Hilfsvariablen gebildet. Im einfachsten Falle wird ein einzelnes, vollständig vorliegendes, kategorisches Merkmal zur Klasseneinteilung verwendet (vgl. Bankhofer, 1995, S. 113). Soll eine stetige Kovariable eingesetzt werden, müssen die Merkmalsausprägungen klassiert werden (Andridge und Little, 2010, S. 42). Erfolgt eine zufällige Zuordnung der Spender zu den Empfängern innerhalb der Imputationsklassen, ähnelt dieses Verfahren der Ziehung einer geschichteten Stichprobe (vgl. Pokropp, 1996, S. 5 f.).

Diese Methode, Ähnlichkeit zu bestimmen, wird auch zur Imputation beim Current Population Survey Outgoing Rotation Group (CPS-ORG) verwendet. Mittels sieben Merkmalen werden hier 11.232 Klassen erzeugt (Bollinger und Hirsch, 2006, S. 487). Joenssen und Müllerleile (2014, S. 9) erzeugen in einer Fallstudie, durch die Verwendung von drei Merkmalen bereits 1.782 Gruppen, von denen sieben Klassen keine Empfänger und 22 Klassen weder Spender noch Empfänger beinhalten. Dies zeigt auch schon Eigenschaften dieser Vorgehensweise auf. Zwar steigt mit jeder verwendeten Auxiliarvariable die Innerklassenähnlichkeit, jedoch steigt auch die Wahrscheinlichkeit, dass Klassen entstehen, in denen wenig oder keine Spender vorhanden sind. Das legt nahe, dass ein Punkt existiert, ab dem die Imputationsgüte negativ beeinflusst wird, wenn ein weiteres Hilfsmerkmal zur Klassenbildung hinzugezogen wird. Dies gilt selbst, wenn das Merkmal einen hohen Erklärungsgehalt für die zu imputierenden Variablen enthält. In der Praxis wurden daher Imputationsmethoden entwickelt, die dynamisch Merkmale, mit deren Hilfe die Adjustment-Cells

konstruiert wurden, außer acht lassen. Mit semipermeabelen Klassengrenzen können Spender über Klassengrenzen hinweg gesucht werden. Hierdurch werden die Impuationsergebnisse dieser hierarchischen Hot-Deck-Verfahren (Kalton und Kasprzyk, 1986, S. 7) stark abhängig von der aufgebauten Hierarchie und wie viele Klassen für die Spendersuche zusammengelegt werden.

Beispiel 3.1: *Klassenbildung mittels Adjustment-Cells*
Für die Datenmatrix des Anhangs A (Fall 2) sollen die Imputationklassen mittels der Adjustment-Cell-Methode bestimmt werden. Zur Bildung der Klassen, welche zur Imputation des Merkmals AGE verwendet werden sollen, werden die Merkmale REGION, SEX und UHRSWORK verwendet. Die Merkmale REGION und SEX sind nominal skaliert, weisen eine geringe Anzahl an verschiedenen Ausprägungen auf und können daher unmittelbar verwendet werden. Das Merkmal UHRSWORK ist metrisch skaliert und grundsätzlich stetig, die Ausprägungen werden klassiert in Vollzeit (UHRSWORK$\geq$ 40) und Nicht-Vollzeit (UHRSWORK$<$ 40). Hierdurch ergeben sich 16 mögliche Imputationsklassen für jede Kombination der Merkmalsausprägungen von REGION, SEX und dem klassierten UHRSWORK.

Die Schwierigkeiten dieser Vorgehensweise zeigen sich bereits bei diesem Beispiel (vgl. Tabelle 3.1). Lediglich drei Klassen enthalten Spender und Empfänger. In drei Imputationsklassen mangelt es an Spendern für die Empfänger. Fünf der Imputationsklassen enthalten zwar Spender, aber keine Empfänger.

	Gemeinsame Merkmalsausprägungen				
Klasse	REGION	SEX	UHRSWORK	Spender	Empfänger
1	S	F	≥ 40	2, 23	11
2	N	M	≥ 40	25	13
3	S	M	≥ 40	1, 9, 14, 18, 20, 22	
4	W	M	≥ 40		16
5	C	F	< 40	17	
6	N	F	< 40		3
7	S	F	< 40	12, 15, 24	
8	W	F	< 40		8, 10
9	C	M	< 40	4, 5	6
10	S	M	< 40	21	
11	W	M	< 40	7, 19	

Tabelle 3.1: Klassenzuordnung mit der Adjustment-Cell-Methode, Anhang A, Fall 2

Entscheidungsbaumverfahren

Eine der Adjustment-Cell-Methode ähnliche Vorgehensweise Imputationsklassen zu konstruieren, stellen die Entscheidungsbaumverfahren dar. Diese Entscheidungsbaumverfahren ähneln grundsätzlich der Adjustment-Cell-Methode, weil beide Ansätze Objekte so klassifizieren, dass sich die Klassenzuordnung grundsätzlich durch „Wenn ..., dann ...“-Regeln nachvollziehen lässt. Wesentliche Unterschiede bestehen jedoch nicht nur in der Auswahl und Reihenfolge der Merkmale, anhand derer der Datensatz getrennt wird, sondern auch, ob alle ausgewählten Merkmale immer verwendet werden. Bei der Konstruktion von Entscheidungsbäumen werden zur Klassenerstellung nur Informationen verwendet, welche auch im Datensatz vorhanden sind. Dagegen werden bei der Adjustment-Cell-Methode grundsätzlich Expertenwissen und weitere Informationen extern zum Datensatz benutzt, um die Merkmale, anhand derer die Objekte in Klassen eingeteilt werden, zu identifizieren. Bereits Brick und Kalton (1996, S. 228) kommentierten, dass manchmal Algorithmen wie CHAID[3] oder CART[4] zur Bestimmung von Imputationsklassen zum Einsatz kommen. Andridge und Little (2010, S. 43) bemerken, dass diese Methoden anwendbar, jedoch wohl nicht sehr verbreitet sind.

Die Erstellung von Entscheidungsbäumen, nicht nur zur Identifizierung von Imputationsklassen, erfolgt grundsätzlich in drei Schritten:

1. Bewertung aller vorhandenen (verbleibenden) Merkmale anhand ihrer gemeinsamen Verteilung mit den zu imputierenden Variablen.

2. Unterteilung des Datensatzes anhand der Merkmalsausprägungen des in Schritt 1 ausgewählten Merkmals.

3. Wiederholung der Schritte 1 und 2 für jeden der aus 2. entstehenden (Teil-) Datensatz, bis ein Abbruchkriterium erfüllt ist.

Zur Feststellung anhand welchem Merkmal die Daten unterteilt werden (2. Schritt), werden im 1. Schritt alle Merkmale, für die mehr als eine Ausprägung vorhanden ist, bewertet. Die Bewertung erfolgt anhand von Kennzahlen die Konzentrationsunterschiede messen. Während der CHAID-Algorithmus (Kass,

[3] Chi-square Automatic Interaction Detectors (Kass, 1980)

[4] Classification and Regression Trees (Breiman et al., 1984)

1980) an dieser Stelle einen χ^2-Unabhängigkeitstest (Pearson, 1900) und der CART-Algorithmus (Breiman et al., 1984) den Gini-Index (Gini, 1912) verwendet, stützen sich der ID3-Algorithmus (Quinlan, 1979; Quinlan, 1983) und dessen Weiterentwicklungen, wie etwa der C4.5-Algorithmus (Quinlan, 1993), auf die Entropiedefinition der Informationstheorie (Shannon und Weaver, 1949, S. 93 ff.). Nach der Bewertung der Merkmale erfolgt eine Aufteilung (split) des Datensatzes gemäß der Merkmalsausprägungen des Merkmals, welches die beste Kennzahl aufweist. Hierbei können die von unterschiedlichen Algorithmen verwendeten Kennzahlen auf unterschiedliche Entscheidungen hindeuten. Zudem ist im Gegensatz zur Adjustment-Cell-Methode auch nicht gesichert, dass in jedem der sich durch die Splits ergebenden Datensätze dasselbe Merkmal im nächsten Iterationsschritt zur Trennung verwendet wird. Abbruchkriterien des Algorithmus können sein:

- Das zu imputierende Merkmal weist für den betrachteten (Teil-) Datensatz lediglich eine Ausprägung auf. In diesem Fall ist der Imputationswert eindeutig.
- Die konstruierte Imputationsklasse enthält keine Spender. In diesem Fall muss eine Fehlerbehandlung stattfinden. Bei C4.5 erfolgt eine Zusammenlegung von Klassen (vgl. Quinlan, 1993, S. 17).
- Eine weitere Trennung würde Imputationsklassen mit zu wenig Spendern erzeugen.
- Die Trennung nach einem weiteren Merkmal verbessert die vom Algorithmus verwendete Kennzahl nur marginal.

Kritisch anzumerken ist, dass Entscheidungsbäume Klassen lediglich in Hinblick auf die Merkmalsausprägungen der Spender konstruieren. Die zweite Forderung von Ford (1983, S. 186) bleibt hierdurch unberücksichtigt. Ob sich dies negativ auf die Ergebnisse einer Imputation auswirkt, wurde in der Literatur noch nicht hinreichend betrachtet. Andererseits werden Probleme, die bei der Adjustment-Cell-Methode auftreten, vermieden. Werden die Abbruchkriterien a priori sinnvoll festgelegt, wird verhindert, dass eine Imputationsklasse zu wenig Spender enthält und daher mit einer anderen Imputationsklasse zusammengelegt werden muss. Eine empirische Betrachtung der Imputationsqualität bei

der Verwendung von Entscheidungsbäumen zur Bildung von Imputationsklassen bieten Creel und Krotki (2006). Verglichen wurden hier die CHAID- und CART-Algorithmen mit dem Ergebnis, dass eine Kombination mit weiteren Verfahren zur Bestimmung von Ähnlichkeiten innerhalb der Imputationsklassen von übergeordneter Bedeutung ist. Eine Betrachtung, ob die Verwendung von Algorithmen zur automatisierten Klassenerstellung einen grundsätzlichen Vorteil bietet, bleibt jedoch aus.

Beispiel 3.2: *Klassenbildung mit einem Entscheidungsbaum*
Abbildung 3.2 zeigt das Ergebnis des C4.5-Algorithmus, welcher auf die Spender des Falls 1 angewendet wurde. Bei Vorhersage der Variable SEX unter Verwendung der Merkmalsausprägungen der Merkmale REGION, AGE, EDUCD und UHRSWORK ist deutlich zu erkennen, dass das Merkmal EDUCD nicht zur Trennung verwendet wird. Des Weiteren werden Objekte, bei denen die Merkmalsausprägungen des Merkmals AGE kleiner-gleich 50 sind, derselben Imputationsklasse zugeordnet. Es findet also keine weitere Unterteilung statt. Werden die Empfänger nun nach dem entstehenden Entscheidungsbaum imputiert, ergibt sich lediglich beim Objekt 11 ein falscher Wert.

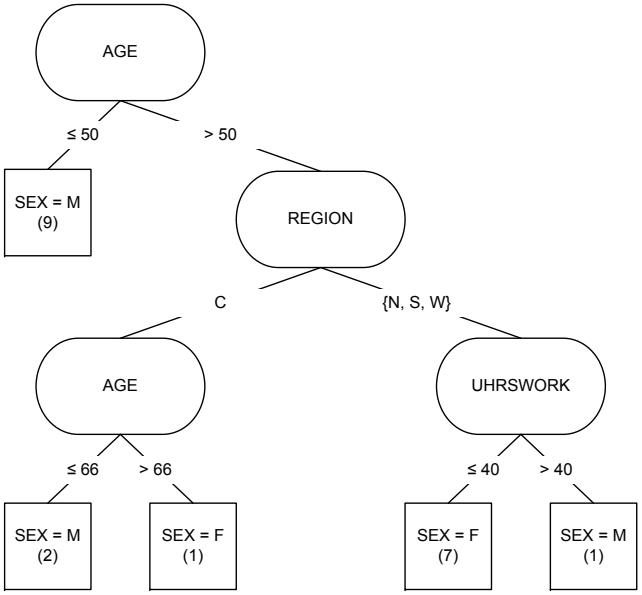


Abbildung 3.2: C4.5 Entscheidungsbaum für Beispiel 3.2

k-Nächste-Nachbarn-Verfahren

Eine früh in der Literatur neben der Adjustment-Cell-Methode diskutierte Vorgehensweise, Imputationsklassen zu bilden, ist das k-Nächste-Nachbarn-Verfahren (k-nearest-neighbor, Fix und Hodges, 1952). Bereits Little (1988, S. 292) und Kaiser (vgl. 1983, S. 523) erwähnten diese Verfahrensvariante im Zusammenhang mit der Hot-Deck-Imputation. Während die Adjustment-Cell-Methode die Spender und Empfänger gemeinsam, basierend auf ihren Werten, in Klassen sortiert, und die Entscheidungsbäume die Empfänger den Klassen zuordnen, die mittels der Spender konstruiert wurden, werden bei dem k-Nächste-Nachbarn-Verfahren die Klassen um die Empfänger herum konstruiert. Im Gegensatz zu den anderen beiden Verfahren werden die Klassen so erstellt, dass diese gezielt nur einen Empfänger und die für die Verdoppelung der Werte infrage kommenden Spender enthalten.

Bei einer k-Nächste-Nachbarn-Klassifikation existieren zwei Parameter, welche die Klassifikation beeinflussen und grundsätzlich frei wählbar sind. Diese sind die Anzahl der Spender, die in eine Klasse aufgenommen werden und das Kriterium, gemäß dessen entschieden wird.

Für den ersten Parameter, die Anzahl der Spender k, kann grundsätzlich jeder Wert zwischen den Extremen von einem Spender und der Anzahl aller vorhandenen Spender d gewählt werden[5]. Es ist daher denkbar, dass ein situatives Optimum für k existiert. Laut Little (1988, S. 292) muss einerseits k, bei der Auswahl eines zufälligen Spenders, groß genug sein, so dass die Verteilung von A^{obs} innerhalb der Imputationsklassen hinreichend der von A entspricht. Anderseits muss k klein genug gewählt werden, so dass die Imputationsklassen nur Spender einer gewissen Ähnlichkeit zum Empfänger enthalten (vgl. Little, 1988, S. 292). Während Kaiser (1989, S. 286) und Colledge et al. (1978, S. 433) einen konkreten Wert von k = 10 annehmen, empfehlen Troyanskaya et al. (2001, S. 522), basierend auf ihren Untersuchungen zu fehlenden Werten bei Gendaten, ein k zwischen 10 und 20. Dahl (2007, S. 5915 f.) hingegen leitet her, dass unter bestimmten Annahmen ein Wert von $k = \lceil\sqrt{d}\rceil$ einer optimalen Klassengröße entspricht, sofern der Spender innerhalb der Klasse rein

[5] Es sei angemerkt, dass sich die Bezeichnung k für die Anzahl der nächsten Nachbarn in der Literatur durchgesetzt hat, und es an dieser Stelle nicht mit dem Index k, der für die Merkmale verwendet wird, gleichzusetzen ist.

zufällig ausgewählt wird und die MAR-Annahme zutrifft. Diesen Wert für eine optimale Klassengröße bestätigen Jönsson und Wohlin (2006, S. 487) mittels durchgeführter Simulationen. Sie empfehlen jedoch, nicht zur nächst größeren ganzen Zahl zu runden, sondern zur nächsten ungeraden Zahl.

Für den zweiten Parameter, das Ähnlichkeitsmaß gemäß dem Spender in die Klassen der Empfänger eingeordnet werden, wird in der Literatur zu Hot-Deck-Verfahren auf zwei Möglichkeiten eingegangen. Zum einen wird die Ähnlichkeit anhand der Objektreihung festgelegt. Hier werden beispielsweise k/2 Spender, die vor und k/2 Spender, die nach dem Empfänger in der Datenmatrix auftauchen, mit in die Imputationsklasse aufgenommen (vgl. Kaiser, 1983, S. 523). Zum anderen werden auch bei der Hot-Deck-Imputation Distanzmaße (vgl. Abschnitt 3.2.1.2) verwendet, um die k-nächsten-Nachbarn zu bestimmen. Für diese Vorgehensweise entwickelten Jhun et al. (2007) eine adaptive Variante k zu bestimmen, welche jedoch bereits elf Jahre zuvor von Schenker und Taylor (1996) entwickelt wurde. In ihrem Verfahren wird k zur Laufzeit, basierend auf einem lokalem Schwellwert, bestimmt (vgl. Schenker und Taylor, 1996, S. 442 ff.; Jhun et al., 2007, S. 1277 f.). Dies ähnelt der Vorgehensweise von Dichte basierten Klassifikationsverfahren (wie etwa dem DBSCAN Algorithmus von Ester et al., 1996) und teilt somit ihre Schwächen (vgl. Schenker und Taylor, 1996, S. 444). Diese Variation der Klassenbildung berücksichtigt zwar besser, welche Spender grundsätzlich für einen Empfänger in Frage kommen, jedoch muss letztendlich anstelle von k ein Schwellwert gewählt werden. Dieser Schwellwert entspricht einer maximalen Distanz, deren Betrag, je nach Merkmalen und Gewichten, die zur Berechnung verwendet wurden, schwerer zu interpretieren ist. Des Weiteren können durch eine falsche Schwellwertwahl leere Klassen auftreten.

Beispiel 3.3: *Klassenbildung mit einem k-Nächste-Nachbarn-Verfahren*
Für die Datenmatrix des Anhangs A (Fall 4) sollen nun die für jeden Empfänger infrage kommenden Spender bestimmt werden. Als Ähnlichkeitsmaß dient die Objektreihung, wobei die Anzahl der Spender k, die für einen Empfänger in Frage kommt, gleich zwei gewählt wird. In Anlehnung an Kaiser (1983, S. 523) wird die eine Hälfte der Spender unmittelbar vor und die andere nach dem Empfänger aus der Datenmatrix entnommen. Durch diese Vorgehensweise ergeben sich jene Imputationsklassen, die in Tabelle 3.2 dargestellt werden.

Imputationsklasse	Empfänger	Mögliche Spender
1	4	3, 6
2	5	3, 6
3	7	6, 9
4	8	6, 9
5	13	12, 14
6	15	14, 16
7	20	19, 21
8	23	22, 24

Tabelle 3.2: Imputationsklassen bei einer k-Nächste-Nachbarn-Klassenbildung unter Nutzung der Objektreihung

Klassenbestimmung unter Nutzung unvollständiger Merkmale

Grundsätzlich wird in der Literatur davon ausgegangen, dass die Klassen für die Spender und Empfänger nur mittels der vollständig vorhandenen Auxiliarvariablen bestimmt werden. Dies würde der Anwendung eines Available-Variable-Verfahrens vor der Klassenbildung entsprechen. Sofern das Ausfallmuster nicht univariat ist, ist es auch möglich, Merkmalsvektoren, bei denen Ausprägungen fehlen, mit einzubeziehen. Während bei der Adjustment-Cell-Methode hier nur eine Option besteht, kann bei den Entscheidungsbäumen unterschieden werden, ob die unvollständigen Merkmalsvektoren zur Klassenidentifikation der Spender oder Empfänger eingesetzt werden sollen.

Zur Konstruktion der Imputationsklassen lässt sich grundsätzlich bei jedem Verfahren das Fehlen einer Merkmalsausprägung als zusätzliche Merkmalsausprägung auffassen. Annahme hierbei ist, dass die Objekte, welche einen fehlenden Wert in diesem Merkmal aufweisen, grundsätzlich bezüglich der zu imputierenden Merkmale anders sind als jene Objekte, die in diesem Merkmal vollständig sind. Dieses Vorgehen lässt sich sowohl in der Adjustment-Cell-Methode als auch bei den Entscheidungsbäumen anwenden. Bei Letzteren führt dieses Vorgehen jedoch zu einer systematischen Bevorzugung von Merkmalen, bei denen viele Werte fehlen, und kann daher nicht als sinnvoll erachtet werden (vgl. Quinlan, 1986, S. 97 f.).

Bei der Klassenzugehörigkeitsidentifikation der Spender, also der eigentlichen Konstruktion des Entscheidungsbaums, ist die Verwendung einer Avail-

able-Case-Analyse möglich. Die erforderlichen Kennzahlen werden dann mittels der paarweise vorhandenen Merkmalsausprägungen berechnet. Diese Vorgehensweise wird durch den CART-Algorithmus zur Berechnung des Gini-Index verwendet (Breiman et al., 1984, S. 142).

Eine alternative Vorgehensweise zur Berücksichtigung von fehlenden Werten bei jenen Objekten, die zur Baumkonstruktion verwendet werden, schlägt Quinlan (1986, S. 98 f.) vor. Zur Berechnung der Kennzahl beim ID3-Algorithmus wird der Anteil einer Klasse an der Kennzahl um einen bestimmten Faktor erhöht. Dieser Faktor enthält jenen Anteil, den die fehlenden Werte in einer bestimmten Klasse ausmachen würden, wenn nach der Zielvariable getrennt wird. Die Spender, die in dem Merkmal, das zur Trennung des Datensatzes verwendet wurde, unvollständig sind, werden im Sinne einer Complete-Case-Analyse vor dem nächsten Iterationsschritt entfernt (Quinlan, 1986, S. 98).

Sind bei den Empfängern auch Merkmale vorhanden, anhand derer eine Klassenzuordnung erfolgen soll, die aber nicht gleichzeitig imputiert werden sollen, kann dies auch mit einbezogen werden. Die Art, wie dies berücksichtigt wird, richtet sich aber danach, wie mit den fehlenden Werten bei der Konstruktion des Entscheidungsbaums umgegangen wurde. Wurde das Fehlen einer Merkmalsausprägung als neue Merkmalsausprägung aufgefasst, muss dies auch bei der Imputationsklasse des Empfängers erfolgen. Sind fehlende Werte anderweitig berücksichtigt worden, muss anders verfahren werden.

Beispielsweise erfolgt beim ID3-Algorithmus die Klassifizierung eines Empfängers mit weiteren fehlenden Werten mit Hilfe eines so genannten „Tokens“. Dieser Token wird, sobald ein Split vorgenommen werden soll, anhand des relativen Anteils der Spender in jeden Ast des Knotens aufgeteilt. Der Empfänger „befindet“ sich daher zu einem gewissen Anteil in jedem folgenden Ast. Diese Anteile dienen dann als Gewichtung der vorhergesagten Imputationsklassen. Falls in einem der folgenden Äste wieder ein Wert fehlt, wird der Token-Anteil erneut aufgeteilt. Am Ende werden die Ergebnisse kombiniert, wodurch sich Wahrscheinlichkeiten der Imputationsklassenzugehörigkeit ergeben. Beim ID3-Algorithmus wird der Empfänger dann jener Imputationsklasse zugeordnet, die in Summe den höchsten Token-Anteil aufweist (Quinlan, 1986, S. 98–99).

Eine weitere Möglichkeit wird beim CART-Algorithmus eingesetzt. Hier wird eine Liste der berechneten Kennzahlen gespeichert (sogenannte Surrogate,

surrogate splits), so dass, sofern eine Merkmalsausprägung fehlt, das nächstbeste Merkmal zur Zuordnungsentscheidung verwendet werden kann (Breiman et al., 1984, S. 40 ff.).

3.2.1.2 Distanzmaße

Eine weitere Möglichkeit, dem Empfänger ähnliche Spender zu identifizieren, ist durch die Berechnung von Distanzen, c_{ij}, zwischen zwei Objekten i und j gegeben. Analog können auch Ähnlichkeitsmaße verwendet werden, da sich diese stets in Distanzen transformieren lassen. Hot-Deck-Verfahren, die sich Distanzen in dieser Art bedienen, sind in der Literatur im Allgemeinen als Nearest-Neighbor-Hot-Deck bekannt (vgl. Kalton und Kasprzyk, 1986, S. 7; Little und Rubin, 2002, S. 69; Andridge und Little, 2010, S. 44). Die benötigten Distanzen werden unter der Verwendung gewisser Merkmalsausprägungen berechnet und können als Zuordnungskosten aufgefasst werden, so dass sich die Auswahl eines geeigneten Spenders als Minimierung dieser Kosten auffassen lässt. Diese Zuordnungskosten müssen auch im Bereich der Hot-Deck-Imputation die Eigenschaften der Nicht-Negativität ($c_{ij} \geq 0$), Symmetrie ($c_{ij} = c_{ji}$) und Identität ($c_{ii} = 0$) (vgl. Jobson, 1992, S. 486) aufweisen.

Bei den Distanzfunktionen lassen sich die Forderungen von Ford (1983, S. 186) für die Imputationsklassen analog formulieren. Demnach müsste die verwendete Distanzfunktion dafür Sorge tragen, dass die Distanzen zwischen den Empfängern und Spendern die Spender-Empfänger-Ähnlichkeit im Hinblick auf die zu imputierenden Merkmale abbildet. Also muss für jeden Empfänger die Distanz zu jenen Spendern kleiner sein, die hinsichtlich der zu imputierenden Merkmale ähnlicher sind. Dies bedeutet, wenn eine Merkmalsausprägung imputiert werden soll und zur Imputation eines Empfängers zwei Spender zur Verfügung stehen, muss bei jenem Spender-Empfänger-Paar die Distanz kleiner sein, bei dem auch die Abweichung im zu imputierenden Merkmal kleiner ist. Da die fehlenden Merkmalsausprägungen der Empfänger aber nicht bekannt sind, ist auch bei der Bildung von Distanzen, analog des Falls der Klassenbildung, eine Überprüfung der Forderung mittels der vorhandenen Daten nicht möglich. Auch hier muss der Forderung mit theoretischen Überlegungen und Expertenwissen begegnet werden, welche in die Merkmalsauswahl und -gewichtung einfließen.

Da eine sehr hohe Vielfalt an möglichen Distanzmaßen existiert, welche beispielsweise Deza und Deza (2006) ausführlichst abhandeln, werden im Folgenden Distanzmaße dargestellt, die im Zusammenhang mit Hot-Deck-Verfahren in der Literatur vorwiegend Erwähnung finden. Weitere Distanzmaße, die seltener in diesem Kontext diskutiert werden, sind im Abschnitt B.3 des Anhangs B zu finden.

Minkowski-Distanz

Die am häufigsten für quantitative Daten in der Praxis angewendeten Distanzfunktionen sind spezielle Ausprägungen der sogenannten Minkowski-Distanz. Dieser auch als gewichtete L_p-Distanz (vgl. Bankhofer, 1995, S. 99) bekannte Distanzindex wird wie folgt auf die vollständig vorhandenen Merkmale berechnet:

$$c_{ij} = \left(\sum_{k=1}^{m-q} \alpha_k \cdot |a_{ik}^{cv} - a_{jk}^{cv}|^p \right)^{\frac{1}{p}} \quad \forall\, i \in D,\, j \in R. \tag{3.1}$$

Die nichtnegativen reellen Zahlen α_k stellen merkmalsspezifische Gewichtungen dar, mittels derer nicht nur der Bezug zu den zu imputierenden Merkmalen hergestellt werden kann, sondern auch Skalenunterschiede ausgeglichen werden können. Alternativ zu dieser Gewichtung können die Merkmale auch, da die einzelnen α_k nichtnegativ sind, a priori mittels $\alpha_k^{\frac{1}{p}}$ reskaliert werden, welches die Anzahl der Multiplikationen von $(m-q)\cdot d\cdot r$ auf $(m-q)\cdot n$ reduziert. Die merkmalsspezifischen Gewichte beziehungsweise die Reskalierungsfaktoren werden in der Missing-Data-Literatur im Wesentlichen auf zwei verschiedene Arten und Weisen gewählt. Zum einen werden Gewichte so gewählt, dass jeder Summand aus (3.1) maximal denselben Beitrag zu c_{ij} zu leisten vermag (vgl. Kozak, 2005, S. 6). Dies kann durch die Wahl $\alpha_k = (\max\limits_i(a_{ik}^{cv}) - \min\limits_i(a_{ik}^{cv}))^{-p}$ erreicht werden, welches garantiert, dass jeder Summand aus (3.1) maximal 1 sein kann. Zum anderen werden die Merkmale einer Z-Transformation unterzogen, um eine Standardabweichung von Eins zu erhalten (vgl. Abdel-Halim und Abdel-Aal, 1999, S. 72; Strike et al., 2001, S. 899). Dies entspricht $\alpha_k = {\sigma_{kk}}^{-p/2}$. Deutlich zu erkennen ist, dass beide Vorgehensweisen lediglich dafür sinnvoll sind, allgemein eine Ähnlichkeit zwischen Spendern und Empfängern zu berechnen. Diese Gewichtungen sind jedoch auf die Abbildung der Spender-Empfänger-Ähnlichkeit im Hinblick auf die zu imputierenden Merkmale nicht zielgerichtet. Somit ist davon auszugehen, dass in der Praxis der Merkmalsaus-

wahl eine höhere Bedeutung zugemessen wird. Denkbar ist es dennoch, dass die Forderungen von Ford (1983, S. 186) in einer zielgerichteten Wahl der Gewichte berücksichtigt werden. Im einfachsten Falle könnte jedes α_k zusätzlich mit dem Betrag der Korrelation zwischen a_{-k}^{cv} und dem Merkmal, welches einer Imputation bedarf, multipliziert werden[6]. Hierdurch würden Merkmale, die einen hohen Bezug zum unvollständigen Merkmal aufweisen, stärker berücksichtigt.

Der Parameter $p \in \mathbb{R}^+$ erlaubt es, steigende merkmalsspezifische Differenzen, je nach Wahl, über- oder unterproportional zu berücksichtigen[7]. Das nachträgliche Ziehen der p-ten Wurzel dient der Normierung und ist für die Auswahl eines Spenders unerheblich, da es lediglich eine monotone Transformation darstellt. Für die Werte $p = 1$, $p = 2$ und $p \to \infty$ haben sich die Namen Manhattan-, Euklidische- und Tschebyscheff-Distanz etabliert. Nähere Erläuterungen dieser sowie Beispielrechnungen können dem Anhang B, Abschnitt B.1 entnommen werden.

Mahalanobis-Distanz

Eine zentrale Entscheidung bei der Berechnung von Distanzen stellt die Auswahl der vollständigen Kovariaten, mittels derer die Distanzen berechnet werden sollen, dar. Problematisch bei der Minkowski-Distanz ist es, dass bei der Auswahl von zwei korrelierten Merkmalen, ein gewisser Anteil der vorhandenen Information mehrfach berücksichtigt wird. Im Extremfall könnten Merkmale ausgewählt werden, bei denen ein Merkmal lediglich eine lineare Transformation eines anderen ist. Hier würde ein Teil der zur Distanzbildung verwendeten Information, auch bei ordnungsgemäßer Gewichtung, ungewollt doppelt berücksichtigt. Diese Problematik kann durch die Anwendung der Mahalanobis-Distanz (Mahalanobis, 1930) vermieden werden. Erstmals von Rubin (1987, S. 158) für die Nutzung im Rahmen der Hot-Deck-Imputation vorgeschlagen,

[6] Sollen mehrere Merkmale gleichzeitig imputiert werden, wäre es denkbar, den Betrag der durchschnittlichen Korrelation zwischen der vollständigen Kovariate und den Merkmalen, die imputiert werden sollen, zu verwenden.

[7] Im Spezialfall $p \geq 1$ sind die Distanzen auch Normen. Sie erfüllen zusätzlich die Dreiecksungleichung $c_{ij} \leq c_{ih} + c_{jh}$.

lässt sich die Mahalanobis-Distanz in diesem Fall wie folgt berechnen[8]:

$$c_{ij} = \sqrt{\left(a_{i-}^{cv} - a_{j-}^{cv}\right)^T \Sigma^{cv^{-1}} \left(a_{i-}^{cv} - a_{j-}^{cv}\right)} \; \forall\, i \in D,\, j \in R. \tag{3.2}$$

Alternativ lässt sich diese Formel auch in Summenform darstellen:

$$c_{ij} = \sqrt{\sum_{k=1}^{m-q} \sum_{l=1}^{m-q} (\Sigma^{cv^{-1}})_{kl} \left(a_{ik}^{cv} - a_{jk}^{cv}\right) \cdot \left(a_{il}^{cv} - a_{jl}^{cv}\right)} \; \forall\, i \in D,\, j \in R, \tag{3.3}$$

wobei $(\Sigma^{cv^{-1}})_{kl}$ das Element aus der k-ten Zeile und der l-ten Spalte von der Inversen der Kovarianzmatrix, welche über die vollständig vorhandenen Variablen berechnet wurde, darstellt. Sind alle Merkmale paarweise unkorreliert ($\sigma_{kl} = 0 \; \forall k \neq l;\; k, l = 1, \ldots, m-q$), reduziert sich die Mahalanobis-Distanz auf eine Euklidische-Distanz mit den merkmalsspezifischen Gewichten $\alpha_k = 1/\sigma_{kk}$. Sind jedoch mindestens zwei Merkmale perfekt korreliert, wird die zugehörige Kovarianzmatrix Σ singulär und kann nicht invertiert werden. Somit ist auch die Berechnung einer Mahalanobis-Distanz nicht möglich. Soll in einem solchen Fall dennoch die Mahalanobis-Distanz verwendet werden, so ist eines der Merkmale von der Berechnung auszuschließen. Eine Beispielrechnung zur Mahalanobis-Distanz mit den Daten aus Anhang A (Fall 3) kann dem Beispiel B.4 (Anhang B, Abschnitt B.2) entnommen werden.

Auch bei der Mahalanobis-Distanz ist es deutlich zu erkennen, dass lediglich eine allgemeine Ähnlichkeit zwischen Spendern und Empfängern berechnet wird. Eine zielgerichtete Abbildung der Spender-Empfänger-Ähnlichkeit im Hinblick auf die zu imputierenden Merkmale wird auch durch diese Distanz nicht garantiert. Um den Forderungen von Ford (1983, S. 186) gerecht zu werden, bietet sich hier nur an, eine entsprechende Auswahl an Kovariaten zu treffen. Grundsätzlich ist zwar eine Modifizierung der Gewichtung denkbar, jedoch müssten die Elemente der invertierten Kovarianzmatrix direkt manipuliert werden. Auswirkungen dessen sind jedoch nicht unbedingt direkt ersichtlich. Eine Modifizierung der Gewichte beeinflusst die positiven Eigenschaften der Mahalanobis-Distanz unter Umständen negativ.

---

[8] Im Allgemeinen wird in der Literatur c_{ij}^2, also die quadrierte Mahalanobis-Distanz, dargestellt (vgl. Mahalanobis, 1936, S. 50). Diese Notation scheint drucktechnisch bedingt, mit dem Ziel, die aufwändige Setzung des Wurzelzeichens zu vermeiden. Beispiele hierfür treten bis in die neuere Zeit auf, wie etwa bei Jobson (1992, S. 487) in der Beschreibung der Euklidischen-Distanz.

Eine einfacher zu interpretierende Form der Mahalanobis-Distanz, die auch weitere Vorteile mit sich bringt, lässt sich unter Verwendung des Konzepts der Hauptkomponenten darstellen[9]. Hierzu seien die Eigenwerte λ_k, die nichttrivialen Lösungen von $\det(\Sigma - \lambda E) = 0$, gegeben wobei E, die entsprechende $k \times k$-Einheitsmatrix darstellt, welche in der Diagonalmatrix der Eigenwerte $\Lambda \in \mathbb{R}^{m \times m}$ zusammengefasst wird. Des Weiteren seien $\lambda_k \geq \lambda_{k+1}\ \forall k = 1, \ldots, m-1$ und Φ_{-k} der zu λ_k gehörige Eigenvektor, der die nichttriviale Lösung von $(\Sigma - \lambda_k E)\,\Phi_{-k} = 0$ ist. So lässt sich, wenn Φ die Matrix der Eigenvektoren $\Phi_{-1}, \ldots, \Phi_{-m}$ darstellt, unter Verwendung einer Eigenwertdekomposition der Matrix[10] Σ^{-1} die Mahalanobis-Distanz als gewichtete Euklidische-Distanz zwischen den Objekten der Hauptkomponenten von A, $B = A\Phi$ schreiben, sofern die hauptkomponentenspezifischen Gewichte $\alpha_k = 1/\lambda_k$ gewählt werden. Dies entspricht der Umformung:

$$\begin{aligned} c_{ij}^2 &= (a_{i-} - a_{j-})^T \Sigma^{-1} (a_{i-} - a_{j-}) \\ &= (a_{i-} - a_{j-})^T \Phi\Lambda^{-1}\Phi^T (a_{i-} - a_{j-}) \\ &= \left(\Phi^T a_{i-} - \Phi^T a_{j-}\right)^T \Lambda^{-1} \left(\Phi^T a_{i-} - \Phi^T a_{j-}\right) \\ &= (b_{i-} - b_{j-})^T \Lambda^{-1} (b_{i-} - b_{j-}) \\ &= \sum_{k=1}^{m} \frac{1}{\lambda_k} \cdot |b_{ik} - b_{jk}|^2 \ \ \forall\, i, j \in N. \end{aligned} \tag{3.4}$$

Übertragen auf den Fall einer unvollständigen Datenmatrix, bei der die Distanzen lediglich über die vollständig vorhandenen Kovariaten berechnet werden sollen, lässt sich die Mahalanobis-Distanz wie folgt darstellen:

$$c_{ij} = \sqrt{\sum_{k=1}^{m-q} \frac{1}{\lambda_k} \cdot |b_{ik}^{cv} - b_{jk}^{cv}|^2}\ \forall\, i \in D,\, j \in R, \tag{3.5}$$

wobei $B^{cv} \in \mathbb{R}^{m-q \times m-q}$ die Hauptkomponenten von A^{cv} bezeichnet.

Ein Vorteil der Darstellung nach Formel (3.5) sind die deutlich erkennbaren Gewichtungen. Die Auswirkungen einer Manipulation dieser sind deutlich leichter zu interpretieren als eine direkte Manipulation der Kovarianzmatrix

[9] Zur vereinfachten Darstellung erfolgt die Entwicklung der neuen Darstellung unter der Verwendung der Datenmatrix A, welche im Anschluss auf den Fall, dass lediglich A^{cv} als Berechnungsgrundlage dient, angepasst wird.

[10] Sofern Σ invertiert werden kann, gilt $\Sigma^{-1} = \Phi\Lambda^{-1}\Phi^T$ da $\Phi^T = \Phi^{-1}$.

oder gar derer Inversen. Denkbar ist daher, dass als hauptkomponentenspezifische Gewichte nicht die Kehrwerte der Hauptkomponentenvarianzen gewählt werden. Möglich wäre eine Selektion von Gewichten, so dass die Distanzen einen Bezug zu den zu imputierenden Merkmalen aufweisen. Im einfachsten Falle könnte dies über die betragsmäßige Korrelation[11] zwischen der entsprechenden Hauptkomponente und dem zu imputierenden quantitativen Merkmal realisiert werden. Soll zusätzlich eine Reskalierung der Hauptkomponenten stattfinden, wäre das Verhältnis zwischen der betragsmäßigen Korrelation und dem Eigenwert als hauptkomponentenspezifisches Gewicht zu wählen. Hierdurch könnten die Differenzen in den Werten der Hauptkomponenten, die einen geringen Zusammenhang mit dem zu imputierenden Merkmal aufweisen, mit einem geringen Gewicht versehen werden. Eine solche Vorgehensweise würde verhindern, dass vorliegende Informationen unbeabsichtigter Weise mehrfach berücksichtigt werden und zugleich den Anforderungen von Ford (1983, S. 186) besser gerecht werden, so dass hierdurch eine Verbesserung von Imputationsergebnissen zu erwarten wäre. Ein umfassendes Beispiel zur Berechnung der Mahalanobis-Distanzen mit Formel (3.5), bei dem zusätzlich gezeigt wird, wie die hauptkomponentenspezifischen Gewichte modifiziert werden können, kann Beispiel B.5 (Anhang B, Abschnitt B.2) entnommen werden.

Ein weiterer Vorteil ist, dass durch die Anwendung von Formel (3.5) die Anzahl der Einzelberechnungen reduziert werden kann. Es ist nicht nur ersichtlich, dass nun die Hauptkomponenten mit ihren entsprechenden Varianzen a priori reskaliert werden, ähnlich der Lösungen bei der L_p-Distanzen, sondern auch, dass durch die vorherige Transformation von A die Anzahl der Gesamtoperationen[12] reduziert wird. Bei einer vollständigen Datenmatrix sind bei Verwendung der ersten Zeile aus (3.4) $2 \cdot m^2 \cdot n \cdot (n-1)$ Operationen notwendig[13], um auf konventionelle Art und Weise alle c_{ij}^2 zu berechnen. Unter der Anwendung

[11] Da die merkmalsspezifischen Gewichte positiv sein müssen, müssen auch die hauptkomponentenspezifischen Gewichte positiv sein. Denkbar wäre auch, das Bestimmtheitsmaß ρ_{kl}^2 anstelle des Betrags der Korrelation zu verwenden.

[12] Unter Operationen ist in diesem Fall von sogenannten standardized floating-point operations (flop) auszugehen (vgl. Ueberhuber, 1997, S. 197).

[13] Additions-, Subtraktions- und Multiplikationsdauer entsprechen je einem flop (vgl. Addison et al., 1993, S. 8; Ueberhuber, 1997, S. 197; Advanced Micro Devices, 2013, S. 167).

der letzten Zeile aus (3.4) sind lediglich $m \cdot n \cdot (n-1) + 2 \cdot m^2 \cdot n$ Operationen erforderlich, sofern die Hauptkomponenten a priori reskaliert werden. Dies bedeutet, dass die einfacher zu interpretierende Form der Mahalanobis-Distanz bei $n = 3$ und $m \geq 2$ sowie bei $n \geq 4$ und $m \geq 1$ weniger Operationen erfordert als die konventionelle Form. Auch das asymptotische Verhalten der zweiten Berechnungsart ist mit $O\left(m^2 \cdot n + n^2 \cdot m\right)$ deutlich besser als $O\left(m^2 \cdot n^2\right)$[14]. Übertragen auf vorliegenden Fall erfordern somit Formel (3.5) $2 \cdot (m-q) \cdot d \cdot r + 2 \cdot (m-q)^2 \cdot (d+r)$ und Formel (3.2) $4 \cdot (m-q)^2 \cdot (d \cdot r)$ Operationen[15]. Im Falle einer Hot-Deck-Imputation mittels der Mahalanobis-Distanz ist auch die Komplexität dieser Vorgehensweisen von der Menge an fehlenden Werten abhängig. Jedoch ist auch hier, wie Tabelle 3.3 entnommen werden kann, der Bereich, in dem die naive Vorgehensweise besser ist, klein.

Anzahl an verwendeter Kovariaten	Anteil fehlender Werte			
	1 %	5%	10%	20%
$m - q = 1$	$n \geq 200$	$n \geq 40$	$n \geq 20$	$n \geq 10$
$m - q \geq 2$	$n \geq 100$	$n \geq 20$	$n \geq 10$	$n \geq 5$

Tabelle 3.3: Wertebereich, in dem die vereinfachte Form weniger Berechnungen erfordert; Anzahl der Empfänger abgerundet

Predictive-Mean-Matching

Eine weitere Methode zur Distanzbildung stellt das Predictive-Mean-Matching dar (Little, 1988, S. 291; Andridge und Little, 2010, S. 44). Dabei wird zunächst eine Regressionsgleichung für jedes Merkmal a_{-k}^{mv}, das fehlende Werte aufweist, aufgestellt. Fehlen bei mehreren Merkmalen Ausprägungen, so muss jedes Merkmal separat behandelt werden. Die benötigten Regressionskoeffizienten werden unter Verwendung aller vollständig vorhandenen Fälle A^{cc} geschätzt. Mittels dieser Regressionsgleichung werden dann Schätzwerte $\hat{a}_{-k}^{mv}$ für alle Objekte, ungeachtet dessen, ob eine Imputation erforderlich ist, für a_{-k}^{mv}

[14] Dies gilt, da die Algorithmen zur Invertierung und Eigenwertdekomposition einer Matrix derselben Komplexitätsklasse angehören (vgl. Trefethen und Bau, 1997, S. 234 ff., 248; Strassen, 1969, S. 355 f.). Hierdurch müssen für den Vergleich der Vorgehensweise lediglich jene Berechnungen nach der Bestimmung von Σ^{-1} beziehungsweise Φ und Λ^{-1} betrachtet werden.

[15] Sofern B a priori reskaliert wird.

berechnet. Als Distanz zwischen den Spendern und Empfängern wird dann die quadrierte Differenz zwischen den Schätzwerten der Spender und Empfänger verwendet. Für eine skalenadäquate Berücksichtigung nominaler Merkmale genügt laut Andridge und Little (2010, S. 44) eine Dummy-Codierung der entsprechenden Merkmale. Beim Predictive-Mean-Matching wird also jedes Merkmal, das fehlende Werte aufweist, separat unter Nutzung des folgenden Distanzmaßes behandelt:

$$c_{ij} = \left(\hat{a}_{ik}^{mv} - \hat{a}_{jk}^{mv}\right)^2 \ \forall\, i \in D,\, j \in R. \tag{3.6}$$

Die Schätzung der Regressionskoeffizenten kann mittels jeglichen, der Skala von a_{-k}^{mv} adäquaten, Regressionsverfahren geschehen. Sollte a_{-k}^{mv} quantitativ sein, so kann beispielsweise eine einfache multiple Regression gemäß dem Kleinste-Quadrate-Prinzip verwendet werden. Wenn a_{-k}^{mv} qualitativ ist, dann könnten die $\hat{a}_{-k}^{mv}$ mittels einer logistischen Regression berechnet werden. Hierbei erfolgt die Koeffizientenschätzung im Kern durch die Maximum-Likelihood-Methode. Die einzelnen Verfahren werden an dieser Stelle nicht weiter dargestellt. Anstelle dessen sei hier auf einschlägige Literatur verwiesen. Eine umfassende Darstellung der multiplen Regression und der Möglichkeiten zum Umgang mit jeglichen Annahmeverletzungen des Modells erfolgt beispielsweise bei Jobson (1991, S. 219 ff.), während die logistische Regression und verwandte Methoden bei Jobson (1992, S. 278 ff.) diskutiert werden.

Beispiel 3.4: *Distanzberechnung beim Predictive-Mean-Matching*
Im folgendem Beispiel soll die Datenmatrix des Anhangs A Fall 3 betrachtet werden. Mittels Predictive-Mean-Matching werden die Distanzen zwischen den ersten beiden Spendern und dem ersten Empfänger, unter der Verwendung der vollständigen Merkmale AGE, TRANTIME und UHRSWORK, berechnet. Da sowohl die Hilfsvariablen und INCTOT quantitativ sind, ist die multiple lineare Regression ein geeignetes Schätzverfahren.

Eine Schätzung der Koeffizienten gemäß dem Prinzip der kleinsten Quadrate, unter Verwendung von A^{cc}, liefert folgendes Modell:

$$\hat{a}_{i9} = -45.112{,}2 + 1.165{,}5 \cdot a_{i4} - 428{,}7 \cdot a_{i7} + 1.205{,}3 \cdot a_{i8}.$$

Mittels dieses Modells werden $\hat{a}_{19} = 23.100{,}75$; $\hat{a}_{29} = 55.506{,}64$ und $\hat{a}_{39} = 22.487{,}34$ geschätzt. Anhand dieser Schätzwerte werden die zwei Spender-Emp-

fänger Distanzen wie folgt berechnet:

$$c_{21} = (55.506{,}64 - 23.100{,}75)^2 = 1.050.141.707$$
$$c_{31} = (22.487{,}34 - 23.100{,}75)^2 = 376.272.$$

Berücksichtigung qualitativer Daten

Grundsätzlich sind für die Berechnung der in Abschnitt 3.2.1.2 aufgeführten Distanzen quantitative Daten erforderlich (vgl. Bankhofer, 1995, S. 99), was die Anwendbarkeit dieser Verfahren für gemischt skalierte Datenmatrizen einschränkt. Sollen Merkmale mit unterschiedlichen Skalen zur Berechnung einer Gesamtdistanz verwendet werden, kann auf zwei verschiedene Arten verfahren werden. Für jedes Merkmal kann eine eigene skalenadäquate Distanz c_{ij}^k berechnet werden. Nachdem diese Einzeldistanzen berechnet wurden, müssen diese dann in einem Folgeschritt zu einem Gesamtdistanzindex aggregiert werden (beispielsweise mittels einer linearen Aggregationsfunktion oder Entscheidungsregeln). Im Falle einer linear homogenen Aggregation werden die für jedes Merkmal separat ermittelten Einzeldistanzen wie folgt aggregiert:

$$c_{ij} = \sum_{k=1}^{m-q} \alpha_k c_{ij}^k \; \forall \, i \in D, \; j \in R. \tag{3.7}$$

Eine andere weitaus gebräuchlichere Methode, Distanzen aus einer gemischten Datenmatrix zu bestimmen, ist es, die qualitativen Merkmale vor der Distanzberechnung neu zu kodieren (vgl. Kim und Curry, 1977, S. 221; Kovar und Whitridge, 1995, S. 411; Peughd und Enders, 2004, S. 547; Graham, 2009, S. 563). Diese Codierung erlaubt eine sinnvolle Anwendung der für quantitative Merkmale vorgesehenen Distanzmaße. Während bei ordinalen Merkmalen stets die Möglichkeit existiert, alle Ausprägungen mittels ihrer Rangplätze zu kodieren, stehen bei nominalen Daten folgende Möglichkeiten zur Verfügung:

> **Dummy-Kodierung:** Bei der Dummy-Kodierung wird der Vektor von Merkmalsausprägungen eines qualitativen Merkmals durch eine Anzahl an Dummy-Variablen ersetzt. Jede Dummy-Variable steht in einem solchen Zusammenhang mit einer der Kategorien des ursprünglichen Merkmalsvektors, so dass, wenn diese Kategorie beobachtet wird, genau diese Dummy-Variable 1 beträgt und alle anderen 0 betragen. Dieses Vorgehen erzeugt genau so viele Dummy-Variablen wie Kategorien im ursprünglichen Merkmal vorhanden sind. Zur Vermeidung von Redundanzen kann

eine Referenzkategorie definiert werden, für welche die entsprechende Dummy-Variable gelöscht wird. Diese Referenzkategorie liegt somit implizit vor, wenn alle verbleibenden Dummy-Variablen den Wert 0 aufweisen (vgl. Fahrmeir et al., 1996b, S. 94).

Effekt-Kodierung: Eine weitere Alternative zur Kodierung qualitativer Merkmale stellt die sogenannte Effekt-Kodierung dar. Die Vorgehensweise ist ähnlich der Dummy-Kodierung, lediglich die Referenzkategorie wird anders behandelt. So entsteht, wie bei der Dummy-Kodierung mit Referenzkategorie, aus einem qualitativen Merkmal eine Indikatorvariable weniger als Kategorien in dem qualitativen Merkmal vorhanden sind. Die Wertezuweisung für die Indikatorvariablen erfolgt analog der Dummy-Kodierung, nur wird, wenn die Referenzkategorie vorliegt, der Wert -1 zugewiesen (vgl. Fahrmeir et al., 1996b, S. 94).

Ein Beispiel zur Anwendung der Codierungsmöglichkeiten, das aufzeigt, welche Vorgehensweise für Distanzberechnungen am sinnvollsten ist, kann Beispiel B.10 (Anhang B, Abschnitt B.4) entnommen werden.

Distanzberechnung unter Nutzung unvollständiger Merkmale

Grundsätzlich wird in der Literatur davon ausgegangen, dass die Distanzen zwischen den Spendern und Empfängern nur mittels der vollständig vorhandenen Auxiliarvariablen berechnet werden. Dies würde der Anwendung eines Available-Variable-Verfahrens vor der Distanzberechnung entsprechen. Es ist jedoch möglich, und sofern das Ausfallmuster nicht univariat ist auch sinnvoll, Merkmalsvektoren, bei denen Ausprägungen fehlen, mit in die Distanzberechnung einzubeziehen. Hierzu ist es denkbar, ein einfacheres Imputationsverfahren vorgelagert oder ein Pairwise-Available-Variable-Verfahren anzuwenden.

Eine Möglichkeit zur Verwendung eines Pairwise-Available-Variable-Verfahren wird in Bankhofer (1995, S. 99 f.) diskutiert. Ausgangspunkt hier sind jene Merkmale, die für den Spender $i \in D$ und den Empfänger $j \in R$ paarweise vorhanden sind. Diese Merkmale werden dann in der Menge $M_{ij} = \{k | v_{ik} = 0 \wedge v_{jk} = 0\}$ zusammengefasst. Da die Anzahl der paarweise vorhandenen Merkmale für jede Spender-Empfänger-Paarung unterschiedlich sein kann, können die einzelnen Distanzen jeweils durch eine unterschiedliche Anzahl von Merkmalsunterschieden entstehen. Dieser Sachverhalt sollte durch eine zusätzliche

Gewichtung jeder Distanz berücksichtigt werden. Im einfachsten Falle kann dies durch die jeweilige Berücksichtigung des Verhältnisses der Anzahl der Merkmale $|M|$ zu der Anzahl der paarweise vorhandenen Merkmalsausprägungen $|M_{ij}|$ geschehen (Bankhofer, 1995, S. 99). Diese Gewichtung nimmt an, dass der Beitrag der nicht paarweise vorhandenen Merkmale zu der Distanzgesamtdistanz im Mittel dem der paarweise vorhandenen Merkmale entspricht. Auf diese Weise ließe sich folgende modifizierte gewichtete L_p-Distanz formulieren:

$$c_{ij} = \left(\frac{|M|}{|M_{ij}|} \sum_{k \in M_{ij}} \alpha_k \cdot |a_{ik} - a_{j_k}|^p \right)^{\frac{1}{p}} \quad \alpha_k \geq 0 \, \forall \, k \in M. \tag{3.8}$$

Beispiel 3.5: *Distanzberechnung mittels der modifizierten L_p-Distanz*
Eine Anwendung der Formel (3.8) auf die ersten drei Objekte des Anhangs A (Fall 1) unter Nutzung der Kovariaten AGE, TRANTIME und INCTOT erfordert zunächst die Bestimmung von $|M|$. $|M| = |\{4, 7, 9\}|$ entspricht der Anzahl der insgesamt für die Distanzberechnung verwendeten Merkmale, und in diesem konkreten Fall, drei. Die Anzahl paarweise vorhandener Merkmalsausprägungen für die betrachteten Objekte sind $|M_{12}| = 2$ und $|M_{32}| = 3$. Die zwei Distanzen werden nun wie folgt berechnet:

$$\begin{aligned} c_{12} &= \sqrt{\frac{3}{2}\left((19-56)^2 + (5-30)^2\right)} &&= 54{,}69 \\ c_{32} &= \sqrt{\frac{3}{3}\left((58-56)^2 + (0-30)^2 + (1.820-11.000)^2\right)} &&= 9.180{,}04, \end{aligned}$$

wobei für die Berechnung $\alpha_k = 1 \; \forall k = 4, 7, 9$ und $p = 2$ zugrunde gelegt wurde.

Die Verwendung von Imputationsverfahren vor der Bestimmung der Spender-Empfänger-Zuordnungskosten, so dass für die Distanzberechnung mehr Merkmale verwendet werden können, stellt eine weitere Möglichkeit dar. Grundsätzlich können an dieser Stelle jegliche anwendbare Imputationsverfahren zum Einsatz kommen. Jedoch erscheint es sinnvoll, sich auf jene Verfahren zu beschränken, die weniger rechentechnisch intensiv und genau sind als das verwendete Hot-Deck-Verfahren an sich. Neben einem rein zufälligen Hot-Deck-Verfahren ist auch eine Lageparameterimputation denkbar. Annahme dieser Vorgehensweise ist, dass die Abweichung bei jenen Merkmalen, die nicht paarweise

bei Spender und Empfänger beobachtet wurden, einer Abweichung zwischen der vorhandenen Merkmalsausprägung und einem mittleren Objekt entspricht. Eine Gegenüberstellung der Auswirkung auf die Imputationsgüte von unterschiedlichen Wegen auch unvollständige Hilfsvariablen zu berücksichtigen, ist bis dato nicht in der Literatur vorhanden.

3.2.1.3 Objektsortierung

Die letzte Methode, um die Wahl eines ähnlichen Spenders für jeden Empfänger zu garantieren, ist die Objektsortierung. Ziel einer solchen Sortierung ist die Erzeugung einer Autokorrelation in den zu imputierenden Variablen (Kalton und Kasprzyk, 1986, S. 7). Es ist hierfür erforderlich, dass mindestens ein weiteres Merkmal vorliegt, das vollständig beobachtet ist und einen hohen Zusammenhang zu den zu imputierenden Variablen aufweist. Im Falle einer stetigen Sortiervariable sollte der Zusammenhang linear sein. Im Falle einer diskreten Sortiervariable sollte sich dieser Zusammenhang durch Gruppenunterschiede widerspiegeln. Diese Sortierung ist zwar besonders relevant für jene Verfahren, bei denen die letzte vorhandene Merkmalsausprägung über die folgenden fehlenden Werte kopiert werden[16], jedoch werden auch die Imputationsergebnisse jener Verfahren beeinflusst, bei denen die Empfängergruppe sequentiell abgearbeitet wird. Es wird in der Literatur zwar kommentiert, dass etwaige Vorteile einer Sortierung wahrscheinlich nicht substantiell sind (vgl. Kalton und Kasprzyk, 1986, S. 7), etwaige Untersuchungen bleiben jedoch bis heute aus.

In der Literatur werden grundsätzlich folgende Sortiermöglichkeiten erwähnt:

- Sortierung nach Objektnummer
- Sortierung nach Zufallszahl
- Sortierung nach:
  - einer Auxiliarvariable
  - mehreren Auxiliarvariablen.

[16] Dieses Verfahren ist auch als das (CPS-)sequential-hot-deck bekannt (Little und Rubin, 2002, S. 68–69).

Sortierung nach Objektnummer

Kalton und Kasprzyk (1986, S. 7) sowie de Waal et al. (vgl. 2011, S. 249–250) diskutieren die Verwendung der ursprünglichen Objektnummerierung als Objektreihenfolge. Diese Verwendung der ursprünglichen Objektnummern entspricht also keiner Neusortierung. Grundsätzlich liegt die Begründung für eine solche Vorgehensweise noch in den frühen Jahren der Anwendung von Hot-Deck-Verfahren. Daten zu aufeinanderfolgenden Fragebögen wurden tendenziell von derselben Person eingegeben, welche dann auch tendenziell die gleichen Fehler machte (vgl. Cox, 1980, S. 721). Gleichzeitig argumentierte man, dass sich Eigenheiten der Datenerhebungsmethode positiv auf die Reihenfolge auswirkten (vgl. Ford, 1983, S. 198). Daten der Befragungen wurden meist nach Region erhoben, bei Telefoninterviews wird nach Vorwahlen gegliedert gefragt und bei persönlichen Befragungen werden Straßenzüge abgearbeitet. Laut Argumentation führt dies dann dazu, dass Objekte, die nebeneinander im Datensatz stehen, auch eine gewisse regionale Nähe aufweisen. Objekte, die einander nahe im Datensatz stehen, sind demnach ähnlich genug für Imputationszwecke (vgl. Ford, 1983, S. 187; Cox, 1980, S. 721; Groves et al., 2004, S. 333). Eine gezielte Neusortierung nach geographischer Zugehörigkeit wird lediglich von Janes (2007) diskutiert. Präsentiert wird ein mögliches Vorgehen für den kanadischen Zensus. Eine Betrachtung der Auswirkung auf die Imputationsgüte bleibt jedoch aus.

Sortierung nach Zufallszahl

Die Sortierung der Daten anhand der Realisierungen einer Zufallszahl resultiert in einer zufälligen Reihenfolge der Objekte. Eine derartige Sortierung wird meist diskutiert, um die Feststellung zu treffen, dass das Sequentielle-Hot-Deck dem Hot-Deck gleicht, bei dem ein Spender zufällig ausgewählt wird (Kalton und Kish, 1984, S. 1921; Kalton und Kasprzyk, 1986, S. 7; Andridge und Little, 2010, S. 46; Joenssen und Bankhofer, 2012, S. 60). Wichtig an diesem Zusammenhang ist nicht etwa, dass für die Durchführung des Sequentiellen-Hot-Decks die Objekte neu sortiert werden sollen, sondern, dass bei der Annahme einer nahezu zufälligen Reihenfolge, für das sequentielle Verfahren Erkenntnisse über das zufällige Verfahren angewandt werden können. Von besonderem Interesse war es, die Verwendung bestimmter Schätzmetho-

den für Parametervarianzen, welche nur für zufällige Verfahren vorhanden sind, zu plausibilisieren (vgl. Oh und Scheuren, 1983, S. 153 ff.; Kalton und Kish, 1984, S. 1922 ff.).

Sortierung nach mehreren Variablen

Soll die Information mehrerer Hilfsvariablen in die Sortierung einfließen, muss eine sogenannte Serpentinen-Sortierung vorgenommen werden (vgl. Grau und Ahmed, 2008, S. 119). In einer Serpentinen-Sortierung wird zunächst der Datensatz nach einer Variable sortiert. Danach erfolgt iterativ eine Sortierung der nächsten Variable innerhalb von Klassen, die durch die Merkmalsausprägungen der vorherigen Variable vorgegeben sind. Eine sinnvolle Sortierung beinhaltet demnach zwei Komponenten. Zum einen muss eine Reihenfolge festgelegt werden, nach der die Variablen zur Sortierung verwendet werden. Da die erste Variable den größten Einfluss auf die Objektreihenfolge ausübt, ist eine sinnvolle Sortiervariablenreihenfolge offensichtlich (vgl. Grau und Ahmed, 2008, S. 119). Die Sortiervariablenreihenfolge sollte der Rangordnung der Zusammenhänge zwischen den Sortiervariablen und den zu imputierenden Variablen entsprechen. Zum anderen muss entschieden werden, wie mit stetigen Variablen zu verfahren ist. Nach einem stetigen Merkmal zu sortieren ist nur sinnvoll, wenn entweder nach einem einzelnen stetigen Merkmal zuletzt sortiert wird oder die Ausprägungen entsprechend in Klassen eingeteilt werden. Wird das Sequentielle-Hot-Deck mit einer Serpentinen-Sortierung zur Imputation verwendet, ähneln die Ergebnisse dem Sequentiellen-Hot-Deck unter Verwendung der Adjustment-Cell-Methode. Größter Unterschied ist, dass bei einer Serpentinen-Sortierung über Klassengrenzen hinweg Imputationswerte übertragen werden könnten. Tabelle 3.4 gibt einen Überblick der Variablen, die in der Literatur zur Sortierung eines Datensatzes vor der Durchführung einer Hot-Deck-Imputation verwendet wurden.

Datensatz	Variablen	Quelle
Census of Construction (Kanada)	Jahresüberschuss	Colledge et al., 1978, S. 433
Longitudinal Daten	Zeitpunkt	Spiess, 2005, S. 253–245
Niederländische Einkommensstruktur	Zufallszahl	de Waal et al., 2011, S. 255
National Household Survey on Drug Abuse	Drogennutzung, Alter, Geschlecht, Rasse	Brittingham, 1998, S. B-6
American Comunity Survey	Region	Joenssen und Müllerleile, 2014, S. 9

Tabelle 3.4: Überblick in der Literatur verwendeter Sortiervariablen

3.2.2 Stochastizität

Die zweite der vier zentralen Eigenschaften der Hot-Deck-Verfahren ist, ob die Auswahl des Spenders für jeden Empfänger deterministisch oder mit einer Zufallskomponente, also stochastisch, erfolgt (Kovar und Whitridge, 1995, S. 408 ff.; Nordholt, 1998, S. 159 f.; Andridge und Little, 2010, S. 41; Joenssen und Bankhofer, 2012, S. 60 f.). Grundsätzlich werden in der Literatur weniger Variationsmöglichkeiten für diese Eigenschaft der Hot-Deck-Verfahren als bei der Bestimmung von Ähnlichkeit diskutiert. Jedoch kann die Entscheidung, ob ein stochastisches Hot-Deck-Verfahren verwendet wird, nicht vollständig unabhängig von der Ähnlichkeitsdefinition geschehen, da bestimmte Kombinationen nur unter bestimmten Ausfallmechanismen zu sinnvollen Imputationsergebnissen führen können. So entstehen unter der rein zufälligen Auswahl eines Spenders, ohne Verwendung eines Ähnlichkeitsmaßes, nur sinnvolle Imputationen, wenn der Ausfallmechanismus MCAR ist (vgl. Andridge und Little, 2010, S. 49). Des Weiteren führt die Verwendung gewisser Ähnlichkeitsmaße oder eine Kombination derer zwangsweise zu einem deterministischen beziehungsweise stochastischen Hot-Deck-Verfahren. Beispielsweise ist ein Hot-Deck-Verfahren, das rein die Spender-Empfänger-Distanzfunktion minimiert, zwangsweise deterministisch.

Beide Alternativen einen Spender zu wählen, weisen situative Vor- und Nachteile auf. Deterministische Hot-Deck-Verfahren zeichnen sich dadurch aus, dass sie mittels desselben unvollständigen Datensatzes einen eindeutigen imputierten Datensatz erzeugen. Dies erlaubt eine Nachvollziehbarkeit, die insbesondere dann wichtig ist, wenn die Datenbereinigung von einer anderen Entität

vorgenommen wird als jene, die den Datensatz darauffolgend analysiert. Hinsichtlich der Entstehungsgeschichte von Hot-Deck-Verfahren ist dies ein großer Vorteil, da Daten der nationalen Statistikämter (insbesondere im Falle des U.S. Bureau of the Census) für andere Parteien veröffentlicht werden. Des Weiteren ist es naheliegend, dass bei einer Bewertung der Spender-Empfänger-Verschiedenheit jener Spender ausgewählt wird, der auch diesem Kriterium bestmöglich entspricht (vgl. Andridge und Little, 2010, S. 43). Stochastische Hot-Deck-Verfahren erzeugen hingegen nicht immer denselben vervollständigten Datensatz, indem sie suboptimale Spender-Empfänger-Paarungen erzeugen. Dies erscheint zunächst kontraintuitiv, jedoch hat diese bewusste Einbringung von Fehlern gewisse Vorteile. Ein Vorteil ist, dass die vervollständigten Daten eine Variabilität aufweisen, die im Mittel besser dem der wahren Daten entspricht (vgl. Schafer und Graham, 2002, S. 167; Haziza, 2007, S. 160). Des Weiteren müssen, wenn multiple Imputation mit Hot-Deck-Verfahren betrieben werden soll, stochastische Hot-Deck-Verfahren verwendet werden, da ansonsten jede der multiplen Imputationen denselben Datensatz liefert (vgl. Allison, 2001, S. 58).

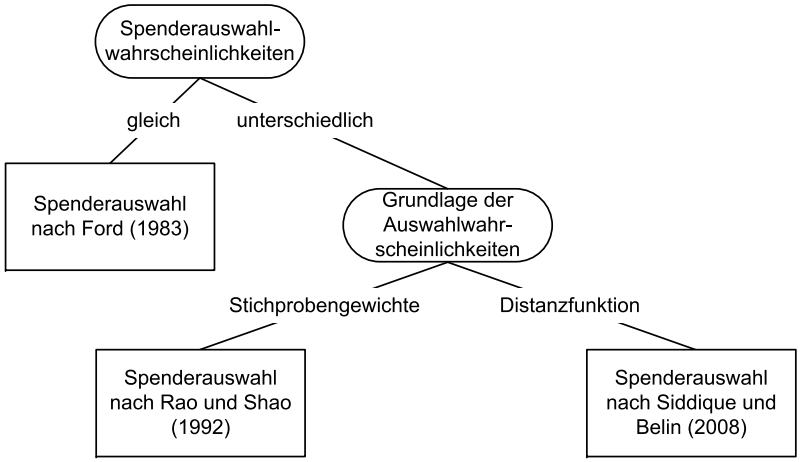


Abbildung 3.3: Möglichkeiten zur Festlegung der Auswahlwahrscheinlichkeiten bei stochastischen Hot-Deck-Verfahren

Zur Auswahl eines zufälligen Spenders existieren in der Literatur drei Möglichkeiten, welche in Abbildung 3.3 dargestellt sind. Zum einen kann jeder Spender einem Empfänger mit gleicher Wahrscheinlichkeit zugeordnet werden (Ford, 1983, S. 196). Zum anderen können die Wahrscheinlichkeiten, mit denen die verschiedenen Spender einem Empfänger zugeordnet werden, unterschiedlich sein. Hierzu haben sich in der Literatur, die sich mit Hot-Deck-Verfahren

beschäftigt, wiederum zwei Möglichkeiten etabliert, die Auswahlwahrscheinlichkeit eines Spenders festzulegen. Die Auswahlwahrscheinlichkeiten orientieren sich entweder an den Werten der Distanzfunktion zwischen den Spendern und Empfängern (Siddique und Belin, 2008, S. 85) oder an den Stichprobengewichten der Objekte (Rao und Shao, 1992, S. 816), wobei im letzteren Fall eine gewichtete Stichprobe vorhanden sein muss.

Die zufällige Spender-Empfänger-Zuordnung, wenngleich das Verfahren interessante Eigenschaften aufweist, ist ein Artefakt der historischen Entwicklung der Hot-Deck-Verfahren. Von Anfang an werden Vermerke in der Literatur gemacht, dass dieses Verfahren und das ursprüngliche sequentielle Hot-Deck-Verfahren äquivalent sind, sofern die Reihenfolge der Objekte in der Datenbasis zufällig ist (vgl. Ford, 1983, S. 192). Unter einer Relaxation des Start-Wert-Problems kann also dieses Hot-Deck-Verfahren als ein zufälliges Ziehen mit Zurücklegen[17] eines Spenders für jeden Empfänger modelliert werden. Der Mehrwert dieser Modellierung besteht darin, dass für bestimmte Ausfallmechanismen die Imputationsvarianz und somit die gesamte Varianz bestimmter Parameter korrekt berechnet werden kann (vgl. Oh und Scheuren, 1983, S. 145). Bei einer gleichen Zuordnungswahrscheinlichkeit jedes Spenders zu einem Empfänger (Oh und Scheuren, 1983, S. 146) stellen sich im Falle keiner Imputationsklassen die Zuordnungswahrscheinlichkeiten wie folgt dar:

$$Pr\left(x_{ij} = 1\right) = \frac{1}{d} \; \forall \, i \in D, \; j \in R. \tag{3.9}$$

Sollte innerhalb von Klassen zugeordnet werden, so gilt diese Formel jeweils innerhalb der erstellten Klassen.

Beispiel 3.6: *Spenderauswahl nach Ford*
Ausgehend von dem Beispiel 3.1, in dem Imputationsklassen mittels der Adjustment-Cell-Methode erstellt wurden, wird im Folgenden innerhalb der Klassen, die Spender und Empfänger beinhalten (Klassen 1, 2 und 9), eine zufällige Spender-Empfänger-Zuordnung gemäß Ford (1983, S. 192) durchgeführt. Für diese Zuordnung ist zunächst eine Ermittlung der Auswahlwahrscheinlichkeiten der Spender für jeden Empfänger erforderlich. Da dies innerhalb der zuvor erstellten Imputationsklassen geschieht, wird jede Imputationsklasse sequentiell

[17] Ein Ziehen mit Zurücklegen wird angenommen, da auch bei dem sequentiellen Verfahren ein Spender für mehrere ihm folgende Empfänger spenden kann.

wie eine separate Datenmatrix behandelt. Zunächst wird die erste Imputationsklasse betrachtet, welche als einzigen Empfänger das Objekt 11 und die zwei Spender, Objekte 2 und 23, enthält. Die einzelnen Auswahlwahrscheinlichkeiten innerhalb Klasse 1 betragen somit

$$Pr\left(x_{i\,11}=1\right)=\frac{1}{2}\;\forall i=2,23.$$

Nach dieser Aufstellung erfolgt eine Zufallsauswahl, basierend auf den ermittelten Wahrscheinlichkeiten. Im Allgemeinen wird hierzu eine auf dem Bereich zwischen 0 und 1 gleichverteilte stetige (Pseudo-) Zufallszahl verwendet. Dieser Bereich wird vorab in aneinander angrenzende Intervalle, deren Länge proportional zu den Auswahlwahrscheinlichkeiten ist, aufgeteilt. Im konkreten Fall wird das Intervall $[0;1]$ in $[0;0{,}5]$ und $(0{,}5;1]$ aufgeteilt. Sollte die generierte Zufallszahl zwischen 0 und 0,5 liegen, ist $x_{2\,11}=1$ zu setzen, und wenn die Zufallszahl im zweiten Bereich liegt, dann wird Objekt 23 dem Objekt 2 als Spender zugewiesen.

Da Klasse 2 nur einen Spender und einen Empfänger enthält, existiert lediglich eine mögliche Zuordnung, und eine Zuweisung kann direkt erfolgen. Die Erstellung der Zuordnungsregeln innerhalb von Klasse 9, bei der die Objekte 4 und 5 für Objekt 6 spenden können, erfolgt analog zu jenen in Klasse 1.

Die grundsätzliche Motivation für die Verwendung von Stichprobengewichten in der Auswahl von einem Spender basiert auf den Aussagen von Cox (1980, S. 721). Demnach würde ein Ignorieren der Stichprobengewichte w bei der Imputation zu einer Verzerrung der Verteilung dieser Gewichte führen, weil ein Spender mit einem geringen Gewicht mit gleicher Wahrscheinlichkeit seine Werte verdoppeln könnte wie ein Spender mit einem hohen Gewicht. Unter Ausnutzung desselben Grundgedankens wie von Oh und Scheuren (1983) bestimmen Rao und Shao (1992) einen Jackknife-Varianzschätzer für das Verfahren von Cox (1980) unter Verwendung der folgenden Wahrscheinlichkeiten, wenn lediglich eine Imputationsklasse vorliegt:

$$Pr\left(x_{ij}=1\right)=\frac{w_j}{\sum\limits_{j\in R}w_j}\;\forall\, i\in D,\; j\in R, \tag{3.10}$$

wobei Auswahlwahrscheinlichkeiten innerhalb der Imputationsklassen analog zu formulieren sind.

Beispiel 3.7: *Spenderauswahl nach Rao und Shao*
Wiederum ausgehend von dem Beispiel 3.1, in dem Imputationsklassen mittels der Adjustment-Cell-Methode erstellt wurden, soll im Folgenden die Vorgehensweise von Rao und Shao (1992) dargestellt werden. Die Vorgehensweise erfolgt größtenteils analog zu jener in Beispiel 3.6, lediglich die Bestimmung der Auswahlwahrscheinlichkeiten für jeden Spender ist anders. Für Klasse 1 ergeben sich, ausgehend davon, dass das Merkmal PERWT die relevanten Gewichte enthält, folgende Wahrscheinlichkeiten:

$$\begin{aligned} Pr\left(x_{2\,11}=1\right) &= \frac{w_2}{w_2+w_{23}} &= \frac{68}{68+91} \\ Pr\left(x_{23\,11}=1\right) &= \frac{w_{23}}{w_2+w_{23}} &= \frac{91}{68+91}. \end{aligned}$$

Der Wertebereich der gleichmäßig stetig verteilten Zufallszahl muss daher auf $[0; 0{,}42767]$ und $(0{,}42767; 1]$ aufgeteilt werden. Sollte die generierte Zufallszahl einen Wert von 0,72118 aufweisen, ist also dem Objekt 11 das Objekt 23 als Spender zuzuordnen. In den Klassen 2 und 9 ist anlog zu verfahren.

Siddique und Belin (2008, S. 85) schlagen eine Strategie zur Nutzung von Distanzen für die zufällige Auswahl von Spendern vor. Die grundlegende Idee dieser Vorgehensweise ist es, die Ähnlichkeitsbeziehungen, die durch Distanzen wiedergegeben werden, zu verwenden, jedoch die Vorteile eines zufälligen Hot-Deck-Verfahrens zu erhalten. Zur Bestimmung der Einzelauswahlwahrscheinlichkeiten für jeden Spender, der für einen Empfänger in Frage kommt, wird folgende Form vorgeschlagen:

$$Pr\left(x_{ij}=1\right) = \frac{\dfrac{1}{c_{ij}}}{\displaystyle\sum_{i\in D}\frac{1}{c_{ij}}},\ \forall\, i \in D,\, j \in R, \tag{3.11}$$

wobei diese Formel garantiert, dass

$$\sum_{i\in D} Pr\left(x_{ij}=1\right) = 1\ \forall\, j \in R.$$

Siddique und Belin (2008, S. 85) verwenden für die c_{ij} ein modifiziertes Predictive-Mean-Matching der Form

$$c_{ij} = \left(|\hat{a}_{ik}^{mv} - \hat{a}_{jk}^{mv}| + \epsilon_j\right)^p\ \forall\, i \in D,\, j \in R \tag{3.12}$$

mit

$$\epsilon_j = \min_j |\hat{a}_{ik}^{mv} - \hat{a}_{jk}^{mv}| \; \forall \, i \in D, \; j \in R \text{ und } \hat{a}_{ik}^{mv} \neq \hat{a}_{jk}^{mv}. \tag{3.13}$$

Der Faktor p dient ähnlich wie bei den Minkowski-Distanzen primär der Skalierung. Allerdings entspricht das von Siddique und Belin (2008) entwickelte Spender-Auswahlverfahren für spezielle Werte von p bestimmten, bereits existierenden, Verfahren. Für $p = 0$ hat jeder Spender dieselbe Auswahlwahrscheinlichkeit und das Verfahren stimmt einem zufälligen Hot-Deck-Verfahren mit gleichen Auswahlwahrscheinlichkeiten für jeden Spender überein. Für $p \to \infty$ liefert das Verfahren dieselben Ergebnisse wie ein deterministisches Predictive-Mean-Matching (vgl. Siddique und Belin, 2008, S. 87).

Die in Formel (3.13) dargestellten ϵ_j dienen als Offset, um Problemen vorzubeugen, die unweigerlich bei Distanzen von Null entstehen würden. Die Autoren kommentieren darauf (vgl. Siddique und Belin, 2008, S. 86), dass die ϵ_j in Fällen, in denen keine zwei Objekte dasselbe Ausfallmuster aufweisen oder identische Prognosewerte für die Merkmale mit den fehlenden Werten aufweisen, vernachlässigt werden können. Da der erstere Fall wohl eher unwahrscheinlich und der zweite Fall a priori nicht bekannt ist, sind die ϵ_j nicht verzichtbar. Allerdings könnte, da lediglich die Division durch Null verhindert werden soll, ein einheitliches Offset von $\epsilon_j = 1 \; \forall j \in R$ dieselbe Wirkung entfalten.

Beispiel 3.8: *Spenderauswahl nach Siddique und Belin*
Im folgendem Beispiel soll die Datenmatrix des Anhangs A (Fall 3) betrachtet werden. Die Bestimmung der Auswahlwahrscheinlichkeiten erfolgt anhand der vollständigen Merkmale AGE, TRANTIME und UHRSWORK sowie der Variable INCTOT, deren Merkmalsausprägungen es für die Objekte $R = \{1; 7; 10; 11; 12; 17; 18\}$ mittels der verbleibenden 18 Objekte zu vervollständigen gilt. Zur Distanzberechnung dient jene von Siddique und Belin (2008) vorgeschlagene Vorgehensweise, so dass das lineare Modell aus Beispiel 3.4

$$\hat{a}_{i1}^{mv} = \hat{a}_{i9} = -45.112{,}2 + 1.165{,}5 \cdot a_{i4} - 428{,}7 \cdot a_{i7} + 1.205{,}3 \cdot a_{i8}$$

verwendet werden kann. Nach Berechnung der einzelnen Prognose muss für jeden Spender das individuelle ϵ_j berechnet werden. Exemplarisch entspricht

das für den ersten Empfänger:

$$\begin{aligned}\epsilon_1 &\approx \min\{|23.100{,}75 - 55.506{,}63|; \ldots; |23.100{,}75 - 28.699{,}93|\} \\ &\approx |23.100{,}75 - 22.487{,}34| \approx 613{,}41.\end{aligned}$$

Hierdurch verändern sich die Distanzen zu den ersten beiden Spendern relativ zu Beispiel 3.4, unter einer Verwendung von $p = 2$, wie folgt:

$$\begin{aligned}c_{21} &\approx (|55.506{,}64 - 23.100{,}75| + 613{,}41)^2 \approx 1.090.273.842 \\ c_{31} &\approx (|22.487{,}34 - 23.100{,}75| + 613{,}41)^2 \approx 1.505.077.\end{aligned}$$

Werden die Distanzen von dem Objekt 1 zu den anderen Spendern berechnet, ergibt sich als Summe von Kehrwerten der Distanzen

$$\sum_{i \in D} \frac{1}{c_{i1}} = \frac{1}{1.090.273.842} + \frac{1}{1.505.077} + \ldots = 9{,}148756 \cdot 10^{-7}.$$

Mit Hilfe dieses Wertes lassen sich für alle Spender die Wahrscheinlichkeiten, mit der sie dem ersten Empfänger zugeordnet werden sollen, berechnen:

$$\begin{aligned}Pr\,(x_{2\,1} = 1) &= \frac{1/1.090.273.842}{9{,}148756 \cdot 10^{-7}} &&= 0{,}001 \\ Pr\,(x_{3\,1} = 1) &= \frac{1/1.505.077}{9{,}148756 \cdot 10^{-7}} &&= 0{,}726 \\ &\vdots &&\vdots \\ Pr\,(x_{25\,1} = 1) &= \frac{1/38.596.239}{9{,}148756 \cdot 10^{-7}} &&= 0{,}028.\end{aligned}$$

In den beiden verbleibenden Schritten müssen nun, wie in den vorherigen Beispielen gezeigt, mittels dieser Wahrscheinlichkeiten das Intervall zwischen $[0; 1]$ aufgeteilt und die Zufallszahl generiert werden. Für die verbleibenden Empfänger sind die Wahrscheinlichkeiten analog zu bestimmen.

3.2.3 Behandlung mehrerer Merkmale

Die dritte Eigenschaft, in der sich Hot-Deck-Verfahren unterscheiden, ist ihr Umgang mit Datenausfall bei mehreren Merkmalen. Entsprechende Varianten unterscheiden sich demnach, wenn das Muster der fehlenden Daten nicht univariat ist (vgl. Abschnitt 2.2). In der Literatur sind drei unterschiedliche Möglichkeiten zu finden, wie hier vorgegangen werden kann:

1. Jedes Merkmal, bei dem fehlende Werte auftreten, wird separat behandelt.

2. Alle Merkmale, bei denen Datenausfall zu verzeichnen ist, werden gleichzeitig behandelt.

3. Der Algorithmus iteriert über die Merkmale.

Hierbei ist es erwähnenswert, dass diese Unterteilung in den ersten zwei Punkten der klassischen Aufteilung von Hot-Deck-Verfahren nach Schnell (1986, S. 109), wie sie auch meist in der deutschsprachigen Literatur zu finden ist, entspricht. Der dritte Punkt ist eine Neuerung die sich basierend auf Weiterentwicklungen der Hot-Deck-Verfahren ergibt. Grundsätzlich ist an dieser Stelle anzumerken, dass für jede der drei Verfahrensvarianten gewisse Vor- und Nachteile in der Literatur, teils implizit, erwähnt und diskutiert werden, die im Folgenden dargelegt werden.

3.2.3.1 Sequentielle und simultane Verfahren

Bei einer sequentiellen Imputation, in der jedes Merkmal mit fehlenden Werten separat betrachtet wird, wird jegliches Ausfallmuster auf eine Anzahl an univariaten Mustern reduziert. Konkret bedeutet dies für Hot-Deck-Verfahren, dass jedes Objekt, welches in dem aktuell behandelten Merkmal einen Wert aufweist, diesen auch spenden kann. Hierdurch entsteht jene Menge an Spendern, die garantiert, dass die Menge an verschiedenen Werten, die verdoppelt werden können, maximal wird (vgl. Sande, 1983, S. 342). Es werden folglich keine Werte unkontrolliert von einer Spende ausgeschlossen[18]. Ein weiterer Vorteil dieser differenzierten Vorgehensweise ist, dass die Hot-Deck-Verfahren auf jedes Merkmal zugeschnitten werden können (vgl. Kalton und Kasprzyk, 1982, S. 28; Marker et al., 2002, S. 333; Andridge und Little, 2010, S. 47 f.). Schließlich bilden in der Praxis Variablen heterogene Inhalte ab, welche von mannigfaltigen Ausfallgründen betroffen sind. Somit wirken verschiedene Ausfallmechanismen gleichzeitig auf die Daten, und es kann davon ausgegangen

[18] Ein kontrollierter Ausschluss von Spendern ist über die geeignete Wahl einer Ähnlichkeitsdefinition zu erreichen.

werden, dass sich der entstehende Mehraufwand einer separaten Behandlung von jedem Merkmal positiv auf die Güte der Imputation auswirkt.

Bei einer simultanen Imputation erfolgt eine Ersetzung sämtlicher fehlender Werte des Empfängers durch die Ausprägungen eines einzigen Spenders (vgl. Bankhofer, 1995, S. 123). Im Falle eines univariaten Ausfallmusters entsprechen die Ergebnisse einer simultanen denen einer sequentiellen Imputation und die Spender- und Empfängermengen sind disjunkt. Weisen mehrere Merkmale fehlende Werte auf, können bei einer simultanen Imputation andere Ergebnisse entstehen und für jeden Empfänger eine andere Menge an Spendern in Frage kommen. So ist es insbesondere bei einem allgemeinen Ausfallmuster möglich, dass ein Objekt einerseits fehlende Werte aufweist und andererseits selbst Werte spendet; die Spender- und Empfängermengen sind nicht mehr notwendigerweise disjunkt.

Welche Objekte für einen gegebenen Empfänger als Spender in Frage kommen, muss mittels der MD-Indikatormatrix bestimmt werden. Ein Objekt $j \in N$ ist immer dann Element der Empfängermenge R, wenn folgende Bedingung gilt:

$$\prod_{k=1}^{m} (1 - v_{jk}) = 0. \tag{3.14}$$

Ein Objekt $i \in N$ kann für ein anderes Objekt $j \in R$ spenden, wenn folgende Bedingung erfüllt ist:

$$\sum_{k=1}^{m} (v_{ik} \cdot v_{jk}) = 0. \tag{3.15}$$

Es ist sofort ersichtlich, dass diese Bedingung symmetrisch ist. Sie zeigt auf, dass zwei Objekte füreinander gegenseitig Spender und Empfänger sein können, wenn deren Ausfallmuster komplementär sind[19]. Denkbar ist es auch, jeden Empfänger als Spender auszuschließen und somit nur Objekte als Spender zuzulassen, bei denen keine Werte fehlen. Ergebnis sind disjunkte Spender- und Empfängermengen. Dies reduziert jedoch weiter die Menge an möglichen Spendern für jeden Empfänger.

Ein häufig angebrachter Vorteil der simultanen Imputation ist, dass durch dieses Vorgehen die gemeinsame Verteilung (Herzog et al., 2007, S. 67), und somit alle multivariaten Eigenschaften (vgl. Kovar und Whitridge, 1995, S.

[19] Die Ausfallmuster zweier Objekte sind komplementär, wenn niemals in beiden Objekten gleichzeitig Werte fehlen.

410), insbesondere die Zusammenhänge der Merkmale untereinander (Longford, 2005, S. 44; Chauvet et al., 2011, S. 460), besser erhalten bleiben. Des Weiteren können auch Inkonsistenzen in den Daten verhindert werden, wenn für alle Imputationswerte derselbe Spender verwendet wird (de Waal et al., 2011, S. 264). Im Allgemeinen erfreut es sich in der Literatur großer Beliebtheit, diese Vorteile der simultanen Imputation zu benennen. So empfehlen Kalton und Kasprzyk (1982, S. 27) sogar, vorhandene Werte zu löschen, um auch diese dann von demselben Spender übernehmen zu können.

Diese vermeintlichen Vorteile der simultanen Imputation sind jedoch kritisch zu bewerten. Zum einen sind diese Aussagen auf die Arbeit von Coder (1978) zurückzuführen, deren mangelnde Verfügbarkeit nicht nur die Überprüfung des Inhalts der Arbeit, sondern auch eine Überprüfung der Zitate in den Sekundärquellen verhindert. Zum anderen sind die Aussagen nicht undifferenziert gültig. Eine simultane Imputation kann nur sinnvoll sein, wenn die Merkmale, die gemeinsam behandelt werden, auch einen inhaltlichen Bezug zueinander aufweisen beziehungsweise mindestens miteinander korrelieren. Nordholt (1998, S. 161) und auch Marker et al. (2002, S. 334) stellen eine noch schärfere Forderung auf. Demnach sollen jene Variablen, die gemeinsam behandelt werden, nicht nur einen starken Bezug zueinander aufweisen, sondern auch die Empfänger tendenziell in denselben Variablen fehlende Werte[20] aufweisen. Werden Variablen, die keinen Bezug zueinander aufweisen, gleichzeitig behandelt, kann folglich angenommen werden, dass durch diese Verfahrensweise Zusammenhänge künstlich geschaffen werden. Selbst das Entstehen von Inkonsistenzen kann im Allgemeinen nicht durch eine gemeinsame Imputation der Variablen verhindert werden (Coutinho und de Waal, 2012, S. 6).

Ein weiterer Nachteil der simultanen Imputation ergibt sich durch die Betrachtung der in Formel (3.15) formulierten Bedingung. Somit kann ein Objekt nur seine Werte spenden, wenn es in allen Merkmalen vollständig ist, in denen der Empfänger fehlende Werte aufweist. Die Anzahl der Objekte, die für einen bestimmten Empfänger spenden kann, ist somit zwangsweise nicht größer als die Anzahl an Objekten, die bei einer sequentiellen Imputation in Frage kommen würde. In der Regel wird also die Menge an Spendern bei der simultanen Imputation größer sein als bei der sequentiellen Imputation. Jedoch ist die

[20] Dieses würde bedeuten, dass die Empfänger ähnliche Ausfallmuster aufweisen sollen.

Anzahl jener für einen Empfänger in Frage kommenden Spender stets bei der sequentiellen Imputation größer als bei der simultanen. Dies kann auch dazu führen, dass unkontrolliert bestimmte Ausprägungen, die in den unvollständigen Merkmalen existieren, von den möglichen Imputationswerten ausgeschlossen werden. Je nach Ausfallmuster und -mechanismus kann sich diese Anzahl an möglichen Werten für einen einzelnen Empfänger auf Null reduzieren (vgl. Beispiel 3.9). Denkbar ist zudem, dass in einem extremen Fall, wie in Beispiel 3.10 dargestellt, die Anzahl der möglichen Spender bei jedem Empfänger unter einem ungünstigen Ausfallmuster null beträgt.

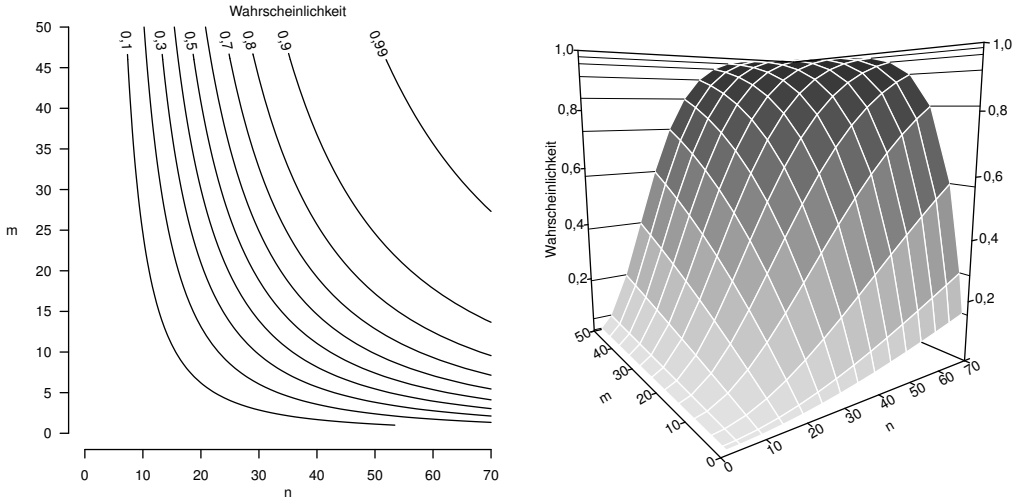
Beispiel 3.9: *Spendermangel beim simultanen Verfahren*
Gegeben sei ein MCAR-Ausfallmechanismus der Form

$$Pr\left(v_{ik}=1\right)=p\;\forall i\in N;k\in 1,\ldots,m,$$

welcher ein allgemeines Ausfallmuster erzeugt, da jeder einzelne Wert der Datenmatrix mit einer konstanten Wahrscheinlichkeit p fehlt. Hierdurch lässt sich unter der Verwendung der Formel von Poincaré und Sylvester die Wahrscheinlichkeit, mit der die Formel (3.15) verletzt wird, bestimmen:

$$\sum_{k=1}^{m}\left(-1\right)^{k-1}\binom{m}{k}\left(\sum_{i=2}^{n}\binom{n}{i}p^{i}\left(1-p\right)^{n-i}\right)^{k}. \tag{3.16}$$

Die Veränderungen der durch Formel (3.16) gegebenen Wahrscheinlichkeiten sind für $n=2,\ldots,100$, $m=1,\ldots,100$ und $p=0{,}01$ in den Abbildungen 3.4a und 3.4b gegeben. Durch eine Betrachtung dieser wird deutlich, dass eine Erhöhung von sowohl n als auch m zu einem Anstieg der Wahrscheinlichkeit führt. Des Weiteren ist ein überproportionaler Anstieg durch gleichzeitige Erhöhung von n und m deutlich ersichtlich. So kann zwar bei einem Datensatz mit $n=10$ Objekten, $m=4$ Merkmalen und $p=0{,}01$ lediglich in ca. 1,69% der Fälle mit einem Empfänger ohne Spender gerechnet werden, jedoch erhöht sich diese Wahrscheinlichkeit bereits auf ca. 12,71% bei 20 Objekten und 8 Merkmalen. Bereits ab $n=120$ und $m=20$ findet sich beinahe immer (ca. 99,97%) mindestens ein Empfänger, der von keinem Spender bedient werden kann.



(a) Höhenlinien (b) Isometrische Ansicht

Abbildung 3.4: Wahrscheinlichkeiten, dass für mindestens einen Empfänger kein Spender existiert, Einzelausfallwahrscheinlichkeit von 1%

Beispiel 3.10: *Unmöglichkeit einer simultanen Imputation*
Die folgend konstruierte MD-Indikatormatrix zeigt beispielhaft, wie eine ungünstige Verteilung der fehlenden Werte in einer Datenmatrix dazu führen kann, dass eine simultane Imputation aller Merkmale unmöglich wird.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

3.2.3.2 Iterative Verfahren

Die Motivation für die letzte Verfahrensvariation bei mehreren Merkmalen mit fehlenden Daten geht auf die Überlegungen von Rubin (1996) zurück. Demnach sollte zur Imputation das Maximum an verfügbaren Informationen verwendet werden (vgl. Rubin, 1996, S. 479). Konkret bedeutet dies, wenn eine Imputation vorgenommen wurde, jene durch die Imputation entstehenden Informationen verwendet werden sollten. Für die sequentiellen Hot-Deck-Verfahren erhöht sich hierdurch die Anzahl der verfügbaren Kovariaten nach jedem behandelten Merkmal um Eins. Dies führt wiederum dazu, dass für Merkmale,

die später behandelt werden, mehr Informationen zur Verfügung stehen.

Eine natürliche Erweiterung (Siddique und Belin, 2008, S. 89) dieser Überlegungen stellt bei Hot-Deck-Verfahren eine erneute Spendersuche nach Vervollständigung des Datensatzes dar. Dies bedeutet, dass alle Merkmale nun unter der durch die vorherigen Imputationen gewonnenen Informationen in der Reihenfolge, in der sie imputiert wurden, reimputiert werden. Diese Vorgehensweise ähnelt einem Gibbs-Sampling (Geman und Geman, 1984), so dass erwartet wird, dass sich die Spenderauswahl mit jeder Iteration bis zur Konvergenz des Verfahrens verbessert (vgl. Siddique und Belin, 2008, S. 89). Vermutet wird zudem, dass die Imputationsergebnisse unabhängig von der Merkmalsimputationsreihenfolge werden (vgl. Siddique und Belin, 2008, S. 89).

Diese ursprünglich von Judkins (1997) unter dem Namen *cyclic n-partition hot-deck* vorgeschlagene Vorgehensweise kann zusammenfassend wie folgt beschrieben werden:

1. Merkmale, die fehlende Werte aufweisen, werden zunächst mittels eines einfachen Impuationsverfahrens vervollständigt.

2. Für jene Ausprägungen, die in einem Merkmal ursprünglich fehlten, werden mit einem Hot-Deck-Verfahren neue Imputationswerte festgelegt. Zur Berechnung der Spender-Empfänger-Ähnlichkeiten werden alle verbleibenden Merkmale der Datenmatrix verwendet.

3. Es erfolgt eine Wiederholung von Schritt zwei für alle Merkmale, die fehlende Werte aufweisen.

4. Schritte zwei und drei werden wiederholt, bis ein Konvergenzkriterium erfüllt ist.

Als einfache Ersetzungsverfahren, die für den ersten Schritt erforderlich sind, kommt jedes Imputationsverfahren in Frage. Jedoch erscheint es sinnvoll, sich an dieser Stelle auf Verfahren zu beschränken, deren Rechenaufwand geringer ist als die darauffolgenden Schritte. Von Siddique und Belin (vgl. 2008, S. 89) wird vorgeschlagen, eine Mittelwertimputation oder ein einfaches zufälliges Hot-Deck-Verfahren zu verwenden. Alternativ könnten beispielsweise auch Distanzen über das von Bankhofer (1995, S. 99 f.) diskutierte Pairwise-Available-Variable-Verfahren berechnet und Startwerte mittels eines Nearest-

Neighbor-Hot-Deck-Verfahrens bestimmt werden. Denkbar ist zudem, dass für die Imputation in jedem Schritt kein sequentielles, sondern ein simultanes Hot-Deck-Verfahren verwendet wird.

Entscheidend für iterative Algorithmen sind das Konvergenzkriterium und die Frage danach, ob dieses erreicht werden kann. Grundsätzlich ist es zweckmäßig, die Veränderung von einem Iterationsschritt zum nächsten in jenen Parameterschätzungen, deren Analyse auch in den nachgelagerten Betrachtungen von Interesse ist, als Konvergenzkriterium zu verwenden. Denkbar ist zudem, dass die Spender-Empfänger-Zuordnung als Konvergenzkriterium verwendet wird. In diesem Fall würden alle denkbaren uni- und multivariaten Parameter gleichzeitig berücksichtigt. Ob jedoch ein *cyclic n-partition hot-deck* in irgendeinem Sinne konvergiert, hängt wohl von den anderen Eigenschaften des konkret angewandten Verfahrens ab. Hierbei erscheint die Stochastizitätseigenschaft des Verfahrens maßgeblich. Bei deterministischen Hot-Deck-Verfahren ist es plausibel, dass Parameterschätzungen oder Spender-Empfänger-Zuordnungen konvergieren. Bei stochastischen Hot-Deck-Verfahren ist es jedoch eher unwahrscheinlich, dass eine Konvergenz hinsichtlich dieser Kriterien erfolgt, da ein gewisses Mindestmaß an Variabilität in den Daten wünschenswert ist. Siddique und Belin (2008) schlagen daher für ihr stochastisches Hot-Deck-Verfahren eine abgewandelte Gelman-Rubin-Statistik (Gelman und Rubin, 1992) vor.

Kritisch zu erwähnen ist, dass die Konvergenz dieses iterativen Verfahrens für keine Hot-Deck-Variante bewiesen wurde. Die Autoren, die sich mit diesem Verfahren befassen, bedienen sich in dieser Hinsicht ausweichender Sprache (vgl. Marker et al., 2002, S. 334–335; Siddique und Belin, 2008, S. 89) und erarbeiten auch keine Empfehlung hinsichtlich einer maximalen Anzahl an Iterationsschritten. Zwar ist es intuitiv plausibel, dass die deterministischen Hot-Deck-Verfahren eine stabile Spender-Empfänger-Zuordnung erreichen, jedoch ist es genauso plausibel, dass entartete Fälle existieren, bei denen zwischen zwei oder mehr Spender-Empfänger-Zuordnungen alterniert wird und keine Konvergenz stattfinden kann. Ferner mangelt es an Belegen, dass, sofern sich die Spender-Empfänger-Zuordnungen verändern, dies tendenziell zu einer Verbesserung der Imputation führt.

3.2.4 Mehrfachverwendung der Spender

Die letzte Eigenschaft, hinsichtlich derer sich Hot-Deck-Verfahren unterscheiden, ist die Bestimmung dessen, wie häufig ein einzelner Spender den Empfängern zugewiesen werden darf. Die Verwendung einer solchen Spenderverwendungshäufigkeitsbegrenzung, im Allgemeinen als Donor-Limit bekannt, ist auf zwei voneinander unabhängige Sachverhalte zurückzuführen. Zum einen geht das Donor-Limit auf die Befürchtung der Anwender von Hot-Deck-Verfahren zurück. Jene waren bei der Anwendung der frühen Hot-Deck-Varianten, welche auf einer rein zufälligen Auswahl der Spender basierten, besorgt, dass aufgrund einer ungünstigen Zufallsauswahl ein Spender „zu häufig“ verwendet wird. Zum anderen geht das Donor-Limit auf die Arbeit von Kalton und Kish (1981) zurück. Sie zeigen, unter Verwendung von Kombinatorik, dass es bei den zufälligen Hot-Deck-Verfahren durch ein Ziehen des Spenderobjekts ohne Zurücklegen zu einer Erhöhung der Schätzpräzision von Verteilungsparametern nach der Imputation kommt (vgl. Kalton und Kish, 1981, S. 147). Eine größtmögliche Einschränkung dessen, wie häufig ein Spender verwendet werden kann, führt demnach zu einer Minimierung der sogenannten Imputationsvarianz. Dies wird in dem folgenden Beispiel exemplarisch dargestellt.

Beispiel 3.11: *Minimierung der Impuationsvarianz*
Die Imputationsvarianz ist bei einem zufälligen Hot-Deck-Verfahren, in dem Spender ohne Zurücklegen ausgewählt werden, kleiner als bei einer Auswahl mit Zurücklegen. Um dies darzustellen, genügt es, ein beliebiges Merkmal mit mindestens zwei fehlenden Werten zu betrachten. Für dieses Merkmal wird dann ein beliebiger Verteilungsparameter für jede mögliche Imputationsvariante berechnet. Zuletzt zeigt sich die kleinere Imputationsvarianz durch die Berechnung der Varianz[21] über die gewählten Verteilungsparameter.

Für die exemplarische Rechnung wird im Folgenden das Merkmal $a_{-1}^T = (1\ 2\ 3\ 4\ 5)$ mit der zugehörigen MD-Indikatormatrix $v_{-1}^T = (0\ 1\ 1\ 0\ 0)$ genutzt. Als Parameter werden exemplarisch der Mittelwert und die Varianz verwendet. Wird das Donor-Limit auf Eins gesetzt, ergeben sich folgende Kombinationen an Imputationswerten mit gleicher Wahrscheinlichkeit, mittels derer sich nach

[21] Da beim Ziehen mit beziehungsweise ohne Zurücklegen alle entstehenden Kombinationen gleiche Wahrscheinlichkeiten haben, genügt eine ungewichtete Varianzberechnung.

einer Imputation die entsprechend dargestellten Mittelwerte und Varianzen ergeben:

Mögliche Imputationswerte						
$a_{2\,1}$	1	1	4	4	5	5
$a_{3\,1}$	4	5	1	5	1	4
$\bar{a}_{\bullet 1}$	3	3,2	3	3,8	3,2	3,8
σ_{11}	3,5	4,2	3,5	2,7	4,2	2,7

Ziehen ohne Zurücklegen.

Ein Donor-Limit von einem Wert größer als Eins führt in diesem Fall zu einem Ziehen von Spendern mit Zurücklegen. Hierdurch ergeben sich folgende Kombinationen an Imputationswerten mit gleicher Wahrscheinlichkeit, mit zugehörigen Mittelwerten und Varianzen:

Mögliche Imputationswerte									
$a_{2\,1}$	1	1	1	4	4	4	5	5	5
$a_{3\,1}$	1	4	5	1	4	5	1	4	5
$\bar{a}'_{\bullet 1}$	2,4	3	3,2	3	3,6	3,8	3,2	3,8	4
σ'_{11}	3,8	3,5	4,2	3,5	2,3	2,7	4,2	2,7	3

Ziehen mit Zurücklegen.

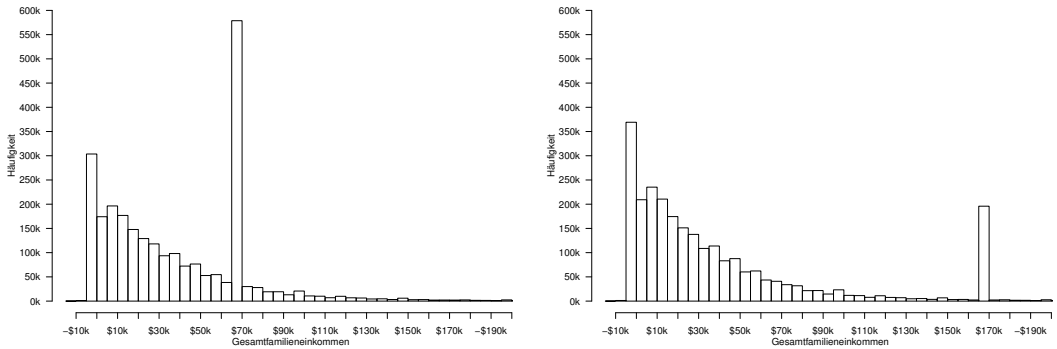
Werden die Varianzen von $\bar{a}_{\bullet 1}$ und $\bar{a}'_{\bullet 1}$[22] beziehungsweise σ_{11} und σ'_{11}[23] verglichen, ist es ersichtlich, dass in beiden Fällen Ziehen ohne Zurücklegen den kleineren Wert liefert.

Für ein Donor-Limit sprechen zudem zwei weitere Gründe. Zum einen wird dadurch das Risiko begrenzt, im Extremfall ausschließlich einen einzigen Spender zu nutzen (vgl. Sande, 1983, S. 345). Zum anderen wird aber auch die Wahrscheinlichkeit reduziert, einen Spender mit extremen Werten zu häufig zu verwenden (vgl. Schnell, 1986, S. 112; Bankhofer, 1995, S. 125; Strike et al., 2001, S. 893), wobei der Begriff „zu häufig“ nie näher definiert wird. Diese Mehrfachverwendung ist in beiden Fällen nicht wünschenswert, da in beiden Fällen ähnliche Effekte wie bei einer Lageparameterimputation auftreten. Wird

[22] 0,1386 versus 0,1506

[23] 0,4506 versus 0,4666

nur ein einziger Spender verwendet, so unterscheidet sich das Hot-Deck-Verfahren von der Lageparameterimputation nur darin, dass nun nicht einmal mehr der Lageparameter gegenüber der Schätzung mittels der vollständigen Objekte erhalten bleibt (vgl. Abbildung 3.5a). Wird ein extremer Wert zu häufig zur Spende verwendet, wird die Variabilität der Daten in ähnlicher Weise beeinflusst. Zudem entsteht eine Verzerrung des Mittelwertes. Diese Effekte sind deutlich in der Abbildung 3.5b zu erkennen.



(a) Ein Spender wird allen Empfängern zugeordnet.

(b) Ein Spender wird 25% aller Empfänger zugeordnet, der Rest zufällig.

Abbildung 3.5: Beispiele für die Risiken der Mehrfachverwendung eines Spenders; Gesamtfamilieneinkommen aus dem American Community Survey (ACS, Ruggles et al., 2010; U.S. Bureau of the Census, 2010), ca. 21% fehlende Werte

Während es beim Basisfall, für jeden Empfänger einen Spender ohne Zurücklegen zu ziehen, eher unwahrscheinlich ist, dass sich diese Risiken manifestieren und daher ein Donor-Limit eher als „beruhigende Maßnahme“ zu betrachten ist (vgl. Sande, 1983, S. 345), sinkt die Wahrscheinlichkeit, dass diese Risiken für manche Variationen der Hot-Deck-Verfahren nicht auftreten. So führen insbesondere zwei Ausprägungen der Eigenschaften von Hot-Deck-Verfahren dazu, dass ein Spender zu häufig verwendet wird. Sollen die fehlenden Werte in allen Merkmalen eines Objekts gemeinsam von nur einem Spender bedient werden, so reduziert sich die Menge an potenziellen Spendern für diesen Empfänger auf jene Objekte der Datenmatrix, die mindestens in allen Merkmalsausprägungen vollständig sind, in denen zugleich auch beim Empfänger Werte fehlen. Wird eine Hot-Deck-Imputation innerhalb von starren Klassenstrukturen vorgenommen, kann dies dazu führen, dass das Verhältnis von Spendern zu Empfängern in manchen oder allen Klassen klein ist, auch,

wenn die Menge an Spendern innerhalb der Datenmatrix an sich hinreichend groß ist (vgl. Andridge und Little, 2010, S. 43). Am problematischsten bei der Erzeugung von Klassen, in denen die Anzahl an Spendern unzureichend ist, ist die beliebte Adjustment-Cell-Methode. Bei diesem Verfahren kann bereits bei einer mäßigen Anzahl an verwendeten Kovariaten eine große Anzahl an Klassen entstehen. Besonders plakativ zeigen dies Bollinger und Hirsch (2006) anhand des Current Population Survey, bei dem bereits die Verwendung von sieben Kovariaten zu der Erstellung von 11.520 Klassen führt, in denen teilweise gar keine Spender vorhanden sind. Auch die Entscheidungsbäume stellen hierbei keine sichere Alternative dar, da diese nur mit Hilfe der Spender erzeugt werden. Letztlich kann eine unerwünscht häufige Nutzung von einzelnen Spendern auch bei den anderen beiden in der Literatur angewandten Möglichkeiten zur Ermittlung von Ähnlichkeitsbeziehungen auftreten. Gerade dann, wenn die Kovariaten, welche einen Erklärungswert für den Ausfall der Werte haben, zum Sortieren verwendet werden, können auf einen einzelnen Spender etliche Empfänger folgen. Darüber hinaus kann auch ein Spender zu häufig zugewiesen werden, wenn dieser mittels einer Distanzfunktion als ähnlichstes Objekt identifiziert wird.

Um diesem Risiko der Spenderübernutzung beizukommen, werden in der Literatur zwei Möglichkeiten diskutiert. Entweder wird ein Donor-Limit explizit gesetzt oder die bereits erfolgte Nutzung eines Spenders für den vorliegenden Datensatz wird mittels eines Strafterms in der Ähnlichkeitsbestimmung berücksichtigt. In dieser bereits 1978 von Colledge et al. (1978, S. 433) für Distanzberechnungen vorgeschlagenen Verfahrensweise werden die adjustierten Spender-Empfänger-Distanzen wie folgt berechnet (vgl. Kalton und Kasprzyk, 1982, S. 24):

$$c'_{ij} = c_{ij} \left(1 + \pi \cdot \sum_{h \in R; h<j} x_{ih} \right) \quad \forall\, i \in D,\, j \in R. \tag{3.17}$$

Hierbei entspricht c_{ij} einer beliebigen Distanz zwischen Spender i und Empfänger j, π einem beliebigen Strafterm für die wiederholte Nutzung des Spenders und $\sum_{h \in R; h<j} x_{ih}$ der Nutzungshäufigkeit des Spenders i für alle bisherig behandelten Empfänger. Zwar wird diese zuerst von Colledge et al. (1978, S. 433) beschriebene Vorgehensweise einige Male in der Literatur erwähnt (vgl. Kalton und Kasprzyk, 1982, S. 24; Kaiser, 1983, S. 523; Kalton und Kasprzyk, 1986,

S. 7; Little, 1988, S. 290; Brick und Kalton, 1996, S. 230; Little und Rubin, 2002, S. 70), jedoch nie konkret die Anwendung aufgezeigt. Daher fehlen nicht nur Empfehlungen zu dem Strafterm π, sondern auch Hinweise auf sinnvolle Werte.

Die Festsetzung eines expliziten Donor-Limits erfreut sich hingegen größerer Beliebtheit. In der Literatur, die sich mit Hot-Deck-Verfahren beschäftigt, findet sich immer wieder die Empfehlung, dass bei Hot-Deck-Verfahren ein Donor-Limit zu verwenden ist (etwa Kalton und Kasprzyk, 1986, S. 7; Brick und Kalton, 1996, S. 229; Nordholt, 1998, S. 179; Strike et al., 2001, S. 893; Little und Rubin, 2002, S. 68; Kim und Fuller, 2004, S. 560; Durrant und Skinner, 2006, S. 27; Durrant, 2009, S. 297 f.; Peláez et al., 2008, S. 811 ff.). Jedoch fehlen auch hier konkrete Empfehlungen oder Untersuchungen, welcher Wert für ein Donor-Limit anzusetzen wäre. Sofern ein Wert diskutiert wird, findet lediglich das stringenteste Donor-Limit Erwähnung. Es wird darauf abgestellt, dass ein Spender nur einmal zu verwenden sei.

Trotz der Tatsache, dass die Verwendung eines Donor-Limits immer wieder mit dem Verweis auf die Vorteile empfohlen wird, führt die Verwendung eines Donor-Limits auch zu gewissen Nachteilen. Sande (1983, S. 345) und Andridge und Little (2010, S. 43) kommentieren, dass eine Einschränkung dessen, wie häufig ein Spender verwendet werden darf, zwangsläufig auch die Möglichkeit, den ähnlichsten Spender auszuwählen, einschränkt und somit die Imputationsqualität negativ beeinflusst. Dies weist auf einen möglichen Verlust an Schätzgenauigkeit hin, welcher von Kalton und Kish (1981) nicht thematisiert wird. Andridge und Little (2010, S. 43) spekulieren daher, dass in Abwägung zwischen Schätzgenauigkeit und -präzision der Verteilungsparameter nach der Imputation, ein optimales Donor-Limit existiert. Des Weiteren kann ein Donor-Limit dazu führen, dass die Imputationsergebnisse abhängig von der Reihenfolge werden, in der die Objekte imputiert werden (vgl. Kovar und Whitridge, 1995, S. 410). Diese Reihenfolge, mit der den Empfängern Spender zugeordnet werden, entspricht jener, mit der die Empfänger in der Datenmatrix auftauchen, und kann durch eine Neusortierung stets verändert werden. Da eine solche Neusortierung der Datenmatrix die Imputationsergebnisse verändern kann, ist diese Eigenschaft im Allgemeinen, insbesondere bei deterministischen Hot-Deck-Methoden, nicht wünschenswert. Deshalb sollte

für deterministische Hot-Deck-Verfahren, bei Verwendung eines Donor-Limits, besser die Gesamtdistanz zwischen allen Empfängern und ihren Spendern minimiert werden.

Aus theoretischer Sicht sprechen somit einige Aspekte für und einige Aspekte gegen die Verwendung eines Donor-Limits. Diese Vor- und Nachteile können jedoch nicht argumentativ aufgelöst werden. Auch eine Meta-Analyse an Literaturquellen scheidet aus, da im Bereich der empirischen Forschung in der Literatur lediglich Studien anzutreffen sind, die Hot-Deck-Verfahren mit anderen Imputationsverfahren vergleichen. Diese Studien betrachten entweder ein zufälliges Ziehen der Spenderobjekte mit Zurücklegen (vgl. Roth und Switzer III, 1995; Barzi und Woodward, 2004; Yenduri und Iyengar, 2007) oder ohne Zurücklegen (vgl. Kaiser, 1983).

Auf Basis dieses Forschungsstands und der weit verbreiteten Nutzung von Hot-Deck-Verfahren mit und ohne Donor-Limit ist es unzufriedenstellend, dass die Notwendigkeit und Konsequenzen des Donor-Limits noch nicht untersucht wurden. Insbesondere fehlen Erkenntnisse darüber, unter welchen Konstellationen das Donor-Limit vorteilhaft beziehungsweise schädlich für die Ergebnisse einer Hot-Deck-Imputation ist. Daher sollen diese Ausführungen als Motivation für die in Kapitel 4 durchgeführten Untersuchungen dienen.

3.3 Das Hot-Deck-Optimierungsproblem

Bei Hot-Deck-Verfahren, die ein Donor-Limit verwenden, kann die Reihenfolge, in der für die Empfänger ein passender Spender gesucht wird, die Spender-Empfänger-Zuordnung beeinflussen. Die schrittweise Wahl des optimalen Spenders für jeden Empfänger führt, gemessen an der Gesamtdistanzsumme, nicht mehr unbedingt zu einer global-optimalen Lösung. Deshalb sollte in diesem Fall besser die Gesamtdistanzsumme, welche sich aus der Spender-Empfänger-Zuordnung ergibt, unter Berücksichtigung der Zuordnungsbeschränkungen minimiert werden. Zu diesem Zweck werden im folgenden Abschnitt 3.3.1 das durch die Einführung des Donor-Limits entstehende ganzzahlige Optimierungsproblem definiert sowie einige Lösungsansätze diskutiert, die eine global-optimale Spender-Empfänger-Zuordnung garantieren. Um mögliche Verbesserungen abzuschätzen, die sich durch solch ein Hot-Deck-Verfahren ergeben,

wird in den darauf folgenden Abschnitten eine vergleichende Simulationsstudie durchgeführt. Hierbei befasst sich Abschnitt 3.3.2 mit der Beschreibung des Studiendesigns und Abschnitt 3.3.3 mit den Ergebnissen der Studie, während in Abschnitt 3.3.4 die Ergebnisse diskutiert werden.

3.3.1 Definition des Problems

Ohne Beschränkung der Allgemeinheit sei anzunehmen, dass Spender und Empfänger zwei disjunkte Mengen an Objekten der Mächtigkeit d beziehungsweise r bilden. Somit kann grundsätzlich jeder Spender für jeden Empfänger spenden[24]. Ferner seien c_{ij} die Elemente der Zuordnungskostenmatrix, die durch Aggregation der Ähnlichkeiten entsteht. Somit lässt sich unter der Verwendung von $dl^{abs} \in \{\lceil {}^{r}/_{d} \rceil, \ldots, r\}$, dem konstanten Donor-Limit, welches für jeden Spender verwendet wird, folgendes ganzzahlige Optimierungsproblem formulieren:

$$\begin{aligned} g\left(x_{ij}\right) &= \sum_{i=1}^{d}\sum_{j=1}^{r} c_{ij}x_{ij} \quad \rightarrow min \\ \sum_{j=1}^{r} x_{ij} &\leq dl^{abs}, \qquad\qquad \forall i = 1,\ldots,d \\ \sum_{i=1}^{d} x_{ij} &= 1, \qquad\qquad \forall j = 1,\ldots,r \\ x_{ij} &\in \{0,1\}, \end{aligned} \tag{3.18}$$

wobei x_{ij} die Zuweisung von Spender i zu Empfänger j angibt und $\lceil \bullet \rceil$ die Aufrundungsfunktion ist. Die zweite Zeile der Formeln aus (3.18) zeigt die Beschränkungen, die durch die Einführung eines Donor-Limits entstehen. Es sind diese Beschränkungen, die eine sequentielle Abarbeitung der Empfänger suboptimal macht. Die dritte Zeile der Formeln aus (3.18) beinhaltet die fundamentale Forderung, dass alle fehlenden Daten durch die Spender-Empfänger-Zuordnung behoben werden müssen. Des Weiteren lässt sich mittels dieser Kombination von Anforderungen und Restriktionen der Wertebereich ermitteln, in dem ein Donor-Limit grundsätzlich variiert werden kann. Die obere Grenze eines möglichen Donor-Limits ist dadurch gegeben, dass ein Spender

[24] Dies ist beispielsweise immer dann der Fall, wenn nur vollständige Objekte als Spender zugelassen werden, ein sequentielles Hot-Deck-Verfahren verwendet wird oder ein univariates Ausfallmuster vorliegt.

maximal jedem Empfänger zugeordnet werden kann. Wird als Donor-Limit r gewählt, so wäre die Zuordnung eines Spenders zu allen Empfängern möglich. Die untere Grenze des Donor-Limits wird durch die Notwendigkeit, dass jedem Empfänger auch ein Spender zugewiesen wird, festgelegt. Eine Festsetzung des Donor-Limits auf einen Wert kleiner als $\lceil {}^{r}/_{d} \rceil$ resultiert darin, dass das Optimierungsproblem nicht lösbar wird.

Eine Optimierung dieser Art, das heißt, die Optimierung der gesamten Spender-Empfänger-Distanzsumme, hat weitere vorteilhafte Eigenschaften, zusätzlich zur Auflösung der obig beschriebenen Reihenfolgenproblematik. Erstens ist es die logische Konsequenz dessen, dass das Angebot an Spendern beschränkt wird. Ohne ein Donor-Limit würden Hot-Deck-Verfahren, die sequentiell den besten Spender jedem Empfänger zuweisen, auch sequenzunabhängig und würden zudem auch die Spender-Empfänger-Distanzsumme minimieren. Zweitens, während weniger Empfänger den bestmöglichen Spender zugewiesen bekommen, sind die Zuweisungen im Mittel besser. Dies sollte sich positiv auf die Imputationswerte kleiner Subpopulationen auswirken, da extreme Abweichungen zwischen den wahren, unbeobachteten Werten und den imputierten Werten weniger häufig werden. Drittens, wenn Spender selten sind oder ein stringenteres Donor-Limit gewählt wird, wird die Auswahl von Ausreißern unumgänglich. Dies ist nahezu paradox, da ein Donor-Limit eigentlich gegen die Verwendung von Ausreißern schützen sollte. In diesem Falle garantiert die Optimierung, dass die Ausreißer unter den Empfängern so verteilt werden, dass die Daten geringst möglich verzerrt werden.

Die grundlegende Struktur des in (3.18) gegebenen ganzzahligen Optimierungsprogramms ist das des klassischen Transportproblems. Obwohl dieses Programm wahrscheinlich eine spezielle Struktur aufweist, kann es durchaus mittels der Algorithmen, die für das Standardproblem entwickelt wurden, gelöst werden. Eine erwähnenswerte Heuristik für die Lösung des klassischen Transportproblems ist die Spaltenminimum-Methode. Wenn die Zuordnungskostenmatrix wie hier definiert ist, so dass die Zeilen die Spender und die Spalten die Empfänger darstellen, ist die Spaltenminimum-Methode äquivalent zu der naiven Vorgehensweise, die heute bei Hot-Deck-Methoden mit einem Donor-Limit verwendet wird. Eine weitere erwähnenswerte Heuristik zur Lösung dieser Zuordnungsproblematik ist Vogels Approximationsmethode (Reinfeld

und Vogel, 1958). Vogels Approximationsmethode ist zwar eine Heuristik, führt aber zu einer reihenfolgenunabhängigen Lösung, da das Auswahlkriterium für eine Zuweisung iterativ über die Distanzmatrix für alle Zeilen und Spalten in jedem Schritt neu berechnet wird (vgl. Domschke, 1995, S. 108). Algorithmen, die garantiert optimale Lösungen liefern, sind beispielsweise die MODI-Methode oder Graphen-basierte Vorgehensweisen. Für eine exhaustive Beschreibung und Diskussion der optimalen Methoden und Heuristiken sei hier auf Domschke (1995) verwiesen.

3.3.2 Studiendesign

Zur Abschätzung, welche Verbesserungen ein Nearest-Neighbor-Hot-Deck-Verfahren, das die gesamte Spender-Empfänger-Distanzsumme unter einem vorgegebenen Donor-Limit minimiert, wird die Imputationsqualität für simulierte Daten verglichen. Die Simulation wird in der Statistiksoftware R in der Version 2.15.2 (R Core Team, 2013) implementiert. Imputationsalgorithmen sind über die Funktion *impute.NN_HD*, als Teil des R-Pakets **HotDeckImputation** (Joenssen, 2013), verfügbar. Die im Folgenden beschriebenen Faktoren entsprechen, mit einiger Vereinfachung, denen in Bankhofer und Joenssen (2014) verwendeten.

Die Daten bestehen strukturell aus zwei gleich großen bivariat-normalverteilten Clustern mit den Zentren $(-1;\ -1)$ bzw. $(1;\ 1)$. Zur Generierung von (Pseudo-) Zufallszahlen, die dieser Verteilung entsprechen, wird die Funktion *rmvnorm* aus dem R-Paket **mvtnorm** (Genz et al., 2013) verwendet. Die Größe der Datenmatrix wird zwischen 50 und 100 Objekten mit drei verschiedenen Innerklassenkorrelationen (0,00; 0,35; 0,75) variiert. Diese Datenmatrizen, welche je 1.000 mal generiert werden, können entweder als eine einzelne Stichprobe oder als eine einzelne Imputationsklasse eines größeren Datensatzes verstanden werden.

Die hieraus resultierenden 6.000 vollständigen Datenmatrizen werden dann drei verschiedenen Ausfallmechanismen ausgesetzt, jeweils mit vier unterschiedlichen Anteilen von fehlenden Werten. Dies wird 1.000 mal wiederholt, so dass 90.000.000 Matrizen mit fehlenden Werten entstehen. Die relativen Häufigkeiten fehlender Werte, bezogen auf die einzige von Ausfall betroffene Variable, sind 10%, 20%, 30%, 40% und 50%. Der erste betrachtete Ausfallmechanismus

ist *MCAR*. Hier wird der vorgegebene Anteil an Werten zufällig gelöscht. Der zweite Ausfallmechanismus, vom Typ *MAR*, löscht Daten, so dass die Menge an fehlenden Werten in einem Cluster immer doppelt so hoch ist wie in dem anderen. Der letzte Ausfallmechanismus ist dem zweiten ähnlich, jedoch ist der Ausfall in dem zweiten Cluster viermal so groß wie in dem ersten.

Die Distanzen zwischen Spendern und Empfängern werden mittels einer Euklidischen Distanz über dem binären Clusterindikator und dem zweiten vollständigen Merkmal berechnet. Die Zuordnung der Spender zu den Empfängern werden dann für alle 90.000.000 Distanzmatrizen durchgeführt, einmal mit der standard-naiven Zuordnungsmethode und einmal mittels der Vogelschen Approximationsmethode[25]. Das Donor-Limit wird auf $dl^{abs} = 1$ festgelegt.

Da Hot-Deck-Verfahren am häufigsten bei Umfragen verwendet werden und hier die Schätzung von Verteilungsparametern meist von Interesse ist (vgl. Little und Rubin, 2002, S. 45), wird die Qualität der Impuation auf Basis von sieben Parametern bewertet. Diese Parameter werden für alle 6.000 vollständigen und alle 180.000.000 imputierten Datenmatrizen, oder ca. 308,4 Gigabyte an Daten, berechnet. Die berechneten Parameter beinhalten vier univariate Parameter und drei multivariate Parameter, welche zwischen der imputierten Variable und der vollständigen Kovariate berechnet werden. Die Parameter sind der Mittelwert ($\bar{a}_{\bullet 1}$), die Standardabweichung ($\sqrt{\sigma_{11}}$), die Schiefe (γ_1), die Kurtosis (β_1), die Mardia-Schiefe ($b_{1,12}$) und Kurtosis ($b_{2,12}$) sowie die Pearson-Korrelation (ρ_{12}). Für jeden dieser Parameter wird die Wurzel aus der mittleren Fehlerquadratsumme (RMSE) zwischen dem wahren Parameter (p_T), auf Basis der vollständigen Datenmatrix, und dem auf Basis der imputierten Datenmatrix geschätzten Parameter (p_I) berechnet, das heißt

$$\text{RMSE} = \sqrt{\frac{1}{1.000.000} \sum_{i=1}^{1.000.000} (p_T - p_I)^2}. \tag{3.19}$$

Die hieraus entstehenden RMSE werden dann gemittelt, um Schätzungen der Haupteffekte zu berechnen.

[25] Zum einen wird auf diese Heuristik zurückgegriffen, weil dieses Verfahren oft nahezu optimale Zuordnungsergebnisse liefert (vgl. Srinivasan und Thompson, 1973, S. 196; Glover et al., 1974, S. 806; Domschke, 1995, S. 105). Zum anderen wird diese Methode verwendet, weil diese in der verwendeten Simulationssoftware implementiert ist.

3.3.3 Ergebnisse

Der folgende Abschnitt beschreibt die Resultate der in Abschnitt 3.3.2 beschriebenen Monte-Carlo-Simulation. Die Werte in den Tabellen 3.5 bis 3.7 sind die Unterschiede in RMSE zwischen den beiden betrachteten Zuordnungsmethoden. Die Differenzen wurden zwischen den RMSE für die konventionelle und der Vogelschen Approximationsmethode gebildet. Daher bedeuten positive Werte, dass Vogels Approximationsmethode besser ist als die naive Vorgehensweise. Werte in den Tabellen sind auf die ersten drei Nachkommastellen beschränkt. Zur Verbesserung der Lesbarkeit werden jene Werte, die auf die ersten drei Nachkommastellen null sind, durch Bindestriche dargestellt.

3.3.3.1 Ergebnisse bei *MCAR*-Ausfall

Die Ergebnisse aus Tabelle 3.5 zeigen, dass die Optimierung der Spender-Empfänger-Distanzsumme niemals unterlegen ist, sofern der Ausfallmechanismus *MCAR* ist. Darüber hinaus ist das auffälligste Ergebnis, dass die Schätzung der univariaten Parameter nur marginal durch die Zuordnungsmethode beeinflusst wird. Weder Mittelwert, Standardabweichung, Schiefe noch Kurtosis der imputierten Variable werden merklich verbessert. Bei den multivariaten Parametern zeigt sich ein anderes Bild. Schätzungen der Pearson-Korrelation sowie der Mardia-Schiefe und Kurtosis sind in jeder Situation deutlich besser, wenn die Daten *MCAR*-Ausfall aufweisen.

		Univariat				Multivariat		
Faktor	Faktorstufen	$\bar{a}_{\bullet 1}$	$\sqrt{\sigma_{11}}$	γ_1	β_1	ρ_{12}	$b_{1,12}$	$b_{2,12}$
Objekt-anzahl	50	−−	−−	−−	−−	0,004	0,086	0,121
	100	−−	−−	−−	−−	0,002	0,067	0,122
Inner-klassen-korrela-tion	0,00	−−	−−	−−	−−	0,001	0,018	0,034
	0,35	−−	−−	−−	−−	0,003	0,051	0,092
	0,70	−−	−−	−−	−−	0,005	0,160	0,238
Anteil fehlender Werte	10%	−−	−−	−−	−−	−−	−−	0,001
	20%	−−	−−	−−	−−	−−	0,001	0,003
	30%	−−	−−	−−	−−	−−	0,004	0,013
	40%	−−	−−	−−	−−	0,001	0,039	0,080
	50%	−−	−−	−−	−−	0,013	0,338	0,510

Tabelle 3.5: RMSE-Differenzen für *MCAR*-Daten, Haupteffekte

Erreichbare Verbesserungen steigen monoton mit der Innerklassenkorrelation und dem Anteil fehlender Werte. Des Weiteren scheinen kleine Datensätze etwas mehr von der globalen Optimierungsmethode zu profitieren.

3.3.3.2 Ergebnisse bei *MAR 1:2*-Ausfall

Tabelle 3.6 zeigt die Ergebnisse für den *MAR 1:2*-Ausfallmechanismus, den Fall, dass ein Cluster doppelt soviel fehlende Werte hat wie der andere. Wieder zeigen die Ergebnisse, dass die Optimierung der Spender-Empfänger-Distanzsumme niemals unterlegen ist und dass primär die Schätzung der multivariaten Parameter verbessert wird. Die Schätzung keiner der univariaten Parameter wird durch die Wahl der Spenderauswahlstrategie substantiell beeinflusst. Bei einem Vergleich der Tabellen 3.5 und 3.6 zeigt sich, dass hier dieselben Effekte vorhanden sind, nur stärker. Die Optimierung wird wichtiger, je mehr Werte fehlen, je stärker die Innerklassenkorrelation ist und um so weniger Objekte die Datenmatrix umfasst.

		Univariat				Multivariat		
Faktor	Faktorstufen	$\bar{a}_{\bullet 1}$	$\sqrt{\sigma_{11}}$	γ_1	β_1	ρ_{12}	$b_{1,12}$	$b_{2,12}$
Objekt-anzahl	50	−−	−−	−−	−−	0,009	0,144	0,160
	100	−−	−−	−−	−−	0,007	0,132	0,150
Inner-klassen-korrela-tion	0,00	−−	−−	−−	−−	0,003	0,038	0,035
	0,35	−−	−−	−−	−−	0,008	0,096	0,105
	0,70	−−	−−	−−	−−	0,012	0,279	0,324
Anteil fehlender Werte	10%	−−	−−	−−	−−	−−	−−	0,001
	20%	−−	−−	−−	−−	−−	0,001	0,004
	30%	−−	−−	−−	−−	−−	0,005	0,015
	40%	−−	−−	−−	−−	0,004	0,166	0,342
	50%	−−	−−	−−	−−	0,034	0,516	0,414

Tabelle 3.6: RMSE-Differenzen für *MAR 1:2*-Daten, Haupteffekte

3.3.3.3 Ergebnisse bei *MAR 1:4*-Ausfall

Bei der Betrachtung von Tabelle 3.7 setzt sich das homogene Bild der Ergebnisse fort. Auch für den dritten Ausfallmechanismus ist Vogels Approximationsmethode niemals schlechter als die naive Vorgehensweise. Beim *MAR 1:4*-

Mechanismus fehlen in einem Cluster viermal soviel Daten wie in dem anderen Cluster, daher stellt dieser Mechanismus eine Verschärfung des *MAR 1:2*-Mechanismus dar. Diese Verschärfung führt zu noch deutlicheren Verbesserungen, wenn Vogels Approximationsmethode verwendet wird. Wieder sind die Vorteile bei der Schätzung der multivariaten Parameter zu finden, wobei die Tendenzen genau denen in den *MCAR*- und *MAR 1:2*-Fällen entsprechen.

		Univariat				Multivariat		
Faktor	Faktorstufen	$\bar{a}_{\bullet 1}$	$\sqrt{\sigma_{11}}$	γ_1	β_1	ρ_{12}	$b_{1,12}$	$b_{2,12}$
Objekt-anzahl	50	−−	−−	−−	−−	0,018	0,149	0,133
	100	−−	−−	−−	−−	0,015	0,150	0,141
Inner-klassen-korrela-tion	0,00	−−	−−	−−	−−	0,008	0,040	0,030
	0,35	−−	−−	−−	−−	0,016	0,107	0,095
	0,70	−−	−−	−−	−−	0,025	0,302	0,286
Anteil fehlender Werte	10%	−−	−−	−−	−−	−−	−−	0,001
	20%	−−	−−	−−	−−	−−	0,002	0,005
	30%	−−	−−	−−	−−	0,001	0,028	0,069
	40%	−−	−−	−−	−−	0,013	0,357	0,120
	50%	−−	−−	−−	−−	0,069	0,360	0,492

Tabelle 3.7: RMSE-Differenzen für *MAR 1:4*-Daten, Haupteffekte

3.3.4 Zusammenfassung

Die Ergebnisse der Monte-Carlo-Simulation zeigen, dass die neue Methode niemals schlechter ist als die konventionelle Zuordnungsmethode, die heutzutage Anwendung findet, wenn ein Donor-Limit eingesetzt wird. Tendenzen, dass die Vorteile stärker werden, sind über alle drei betrachteten Ausfallmechanismen konsistent. Die Vorteile steigen immer, je weniger Spender in den Daten vorhanden sind und je stärker die Korrelation der Variablen innerhalb der Klassen ist. Zudem werden Vorteile des neuen Ansatzes kleiner, je größer die Datenmatrix und je weniger schwerwiegend der Ausfallmechanismus ist.

Am interessantesten ist wahrscheinlich, dass nur die Schätzung multivariater Parameter substantiell durch die Optimierung verbessert wird. Das deutet darauf hin, dass der primäre Wert des erhöhten Aufwands, die gesamten Spender-Empfänger-Zuordnungskosten zu minimieren, in der besseren Erhal-

tung von multivariaten Eigenschaften der Daten liegt. Dies ist besonders dann wichtig, wenn die Daten, nach Imputation, mittels multivariater Verfahren, zum Beispiel multiple Regression, ausgewertet werden sollen.

Diese Ergebnisse lassen sich für eine breite Menge an Situationen und Anwendungen verallgemeinern. Nicht nur wurden die wichtigsten Faktoren in der Simulationsstudie berücksichtigt, sondern es wurde auch nur eine kleine Menge an Annahmen getroffenen. Trotzdem existieren Einschränkungen der Verallgemeinerbarkeit der Ergebnisse. Erstens decken zwar die hier betrachteten Ausfallmechanismen, bei denen zwei Untergruppen unterschiedliche Ausfallwahrscheinlichkeiten haben, viele Situationen gut ab, aber es sind weitaus mehr MAR-Mechanismen denkbar. Zweitens kann in der Praxis eine Hot-Deck-Imputation mit einem Donor-Limit von Eins nicht durchführbar sein. Insbesondere durch die in der Praxis gängigen Methoden zur Konstruktion von Imputationsklassen und wenn bei mehreren Variablen fehlende Werte auftreten, kann die Menge an Empfängern leicht größer werden als die Menge an Spendern. Eine Überprüfung weniger stringenter Donor-Limits sollte weitere interessante Ergebnisse liefern. Drittens wurde lediglich eine Datenstruktur betrachtet. Datenstrukturen, die in der Realität existieren, sind komplexer, und daher (insbesondere wegen der Präsenz von Ausreißern) reagieren sie sensibler auf die Zuordnung schlechter Spender. Viertens wurden nur die Ergebnisse der konventionellen Methode mit jenen einer Heuristik, die reihenfolgeunabhängige Ergebnisse erzeugt, verglichen. Zukünftige Arbeiten könnten in einem weiteren Schritt nicht nur die zwei Heuristiken, sondern auch eine optimale Methode im Hinblick auf Imputationsgüte, aber auch im Hinblick auf Berechnungszeit und -komplexität, vergleichen. Des Weiteren könnten die spezielle Struktur des ganzzahligen Optimierungsproblems ausgenutzt und spezialisierte, effiziente Algorithmen zur Lösung dieses speziellen Problems entwickelt werden.

Kapitel 4

Auswirkungen des Donor-Limits

Durch die Verdoppelungseigenschaft der Hot-Deck-Verfahren stellt sich das bereits erläuterte Problem, dass ein Spenderobjekt wiederholt zur Imputation ausgewählt werden kann. Hierdurch ergeben sich die Risiken, dass zum einen ein Spender „zu häufig“ oder gar für alle Empfänger verwendet werden könnte oder zum anderen ein Ausreißer als Spender gewählt wird. Die Wahrscheinlichkeit, dass sich diese Risiken manifestieren und somit sub-optimale Imputationsergebnisse realisiert werden, wird durch die Art und Weise, wie in der Praxis Imputationsklassen definiert werden, erhöht. Die Beantwortung der Frage, ob und unter welchen Umständen ein Donor-Limit bessere Imputationsergebnisse verspricht, ist somit nicht nur von theoretischem Interesse, sondern auch für die Praxis von besonderer Bedeutung.

Nachdem der Forschungsstand zum Donor-Limit bereits in Abschnitt 3.2.4 dargestellt wurde, wird in den folgenden Abschnitten 4.1 und 4.2 der Sachverhalt mittels Simulation näher betrachtet. Es bietet sich an, diese Untersuchungen in eine Vor- und Nachuntersuchung aufzuteilen. In der Voruntersuchung (Abschnitt 4.1) werden zunächst einige grundlegende Fragen beantwortet. Mit Hilfe der Antworten auf die Forschungsfragen und weiterer Erkenntnisse aus dieser Voruntersuchung wird in Abschnitt 4.2 eine weitere Untersuchung durchgeführt. Diese liefert mittels der Ergebnisse einer umfassenden Simulationsstudie zur Verwendung eines Donor-Limits belastbare Erkenntnisse.

4.1 Voruntersuchung

Im Hinblick auf eine Untersuchung der Auswirkungen eines Donor-Limits im Rahmen einer Hot-Deck-Imputation lassen sich insgesamt die folgenden vier konkreten Forschungsfragen ableiten, die mit der Simulationsstudie beantwortet werden sollen:

1. Ist grundsätzlich ein Donor-Limit im Rahmen einer Hot-Deck-Imputation sinnvoll beziehungsweise notwendig?

2. Von welchen Gegebenheiten des vorliegenden Datenmaterials hängt eine notwendige Beschränkung ab?

3. Ist die Vorteilhaftigkeit eines Donor-Limits vom verwendeten Hot-Deck-Verfahren abhängig?

4. Können in den jeweils betrachteten Fällen weitere Empfehlungen hinsichtlich des Donor-Limits gegeben werden?

Wie anhand der aufgezeigten Fragestellungen zu sehen ist, soll zunächst untersucht werden, ob ein Donor-Limit überhaupt Auswirkungen auf die anschließende Analyse der vervollständigten Daten hat. Falls dies gezeigt werden kann, sind anschließend die Zusammenhänge mit den vorliegenden Daten und dem verwendeten Hot-Deck-Verfahren zu untersuchen. Darüber hinaus ist dann noch zu klären, welche Empfehlungen bezüglich des konkreten Donor-Limits abgeleitet werden können.

Diesen Fragestellungen soll im Rahmen der nächsten Abschnitte nachgegangen werden. Dazu wird in Abschnitt 4.1.1 zunächst das Studiendesign vorgestellt. Im anschließendem Abschnitt 4.1.2 werden die Ergebnisse der Simulation präsentiert. Der letzte Abschnitt fasst die Ergebnisse der Studie zusammen.

4.1.1 Studiendesign

Im Rahmen der Simulationsstudie soll nun untersucht werden, welche Auswirkungen ein Donor-Limit bei einer Hot-Deck-Imputation auf die vervollständigte Datenmatrix hat. Dazu erfolgt im nächsten Abschnitt zunächst eine Darstellung jener Faktoren, die einen Einfluss auf die Untersuchungsergebnisse

erwarten lassen und daher in der Simulationsstudie variiert werden. Im Anschluss daran erfolgt eine Darstellung der gewählten Gütekriterien, welche zur Beurteilung der Simulationsergebnisse verwendet werden. Abschließend wird noch auf die konkrete Durchführung der Simulationsstudie eingegangen.

4.1.1.1 Einflussfaktoren

Auf Basis der Sichtung thematisch ähnlich gelagerter Studien (Roth und Switzer III, 1995; Roth et al., 1999; Strike et al., 2001) sowie ergänzender Überlegungen können eine Reihe möglicher Faktoren identifiziert werden, die im Zusammenhang mit der Notwendigkeit eines Donor-Limits im Rahmen einer Hot-Deck-Imputation gegebenenfalls eine Rolle spielen. Diese Faktoren werden folglich in der Simulationsstudie variiert, um dadurch entsprechende Einflüsse analysieren zu können. Im Einzelnen sind folgende Einflussfaktoren einschließlich der für diese Studie festgelegten Ausprägungen zu nennen:

Dimension der Datenmatrix: Ausgehend von n Objekten und m Merkmalen werden vier unterschiedlich dimensionierte $(n \times m)$-Datenmatrizen mit den Dimensionen (100×9), (350×9), (500×9) und (1750×9) betrachtet. Die betrachteten Größenordnungen resultieren aus den nachfolgend noch aufgeführten Vorgaben zu den betrachteten Merkmalen sowie den Imputationsklassen.

Skalierung der Merkmale: Es werden gemischt skalierte Datenmatrizen mit jeweils drei nominalen, ordinalen und quantitativen Merkmalen betrachtet. Als nominal skalierte Merkmale werden ausnahmslos dichotome Merkmale berücksichtigt, da grundsätzlich auch ein nominal polytomes Merkmal durch eine entsprechende Anzahl dichotomer Merkmale dargestellt werden kann. Bei den ordinal skalierten Merkmalen werden die beiden Fälle von 5 und 7 unterschiedlichen Ausprägungen unterstellt.

Anzahl der Imputationsklassen: Die Imputationsklassen werden vorab bereits festgelegt, das heißt, es werden Daten entsprechend der jeweils festgelegten Klassen generiert. Dazu wird mit 2 eine eher geringe Anzahl und mit 7 eine höhere Anzahl von Klassen festgelegt.

Anzahl der Objekte je Imputationsklasse: Für die bereits vorab festgelegten Imputationsklassen sollen die Fälle einer geringen und einer eher höheren Anzahl von Objekten je Klasse unterschieden werden. Dabei wird zwischen einer Anzahl von 50 und 250 Objekten je Klasse unterschieden.

Klassenstruktur: Für die Imputationsklassen sollen des Weiteren die Fälle einer schwachen sowie einer relativ starken Klassenstruktur unterschieden werden. Von einer starken Klassenstruktur wird ausgegangen, wenn sich die Klassen nur bis maximal 5% überlappen und die Merkmale innerhalb der Klassen eine paarweise Korrelation von mindestens 0,5 aufweisen. Demgegenüber wird für den Fall einer schwachen Klassenstruktur unterstellt, dass die Klassen eine Überlappung von mindestens 30% besitzen und die Merkmale paarweise unkorreliert sind.

Anteil fehlender Daten: Hier werden die Fälle 5%, 10% und 20% fehlender Werte, bezogen auf die gesamte Datenmatrix, unterschieden.

Ausfallmechanismus: Es werden die zwei unsystematischen Ausfallmechanismen MCAR und MAR sowie der Fall NMAR unterschieden.

Hot-Deck-Verfahren: In der Simulationsstudie werden sechs Hot-Deck-Verfahren betrachtet. Diese Verfahren, „SeqZ“, „SeqD“, „SeqDL“, „SimZ“, „SimD“ und „SimDL“, sind nach den Eigenschaften benannt, welche sie aufweisen. Die Präfixe deuten an, ob es sich um ein sequentielles („Seq“) oder simultanes („Sim“) Hot-Deck-Verfahren handelt. Die Postfixe geben an, ob innerhalb der Imputationsklassen die Spender den Empfängern zufällig („Z“) oder distanzbasiert („D“ beziehungsweise „DL“) zugeordnet werden. Bei der ersten Distanzbestimmung wird eine linearhomogene Aggregation auf Basis der paarweise vorhandenen Daten durchgeführt (Verfahrensbezeichnungen „SeqD“ und „SimD“). Bei der zweiten Distanzbestimmung erfolgt eine linearhomogene Aggregation nach einer zuvor erfolgten Imputation mit Hilfe des Lageparameters (Verfahrensbezeichnungen „SeqDL“ und „SimDL“). Zur Bestimmung der merkmalsweisen Distanzen werden bei den quantitativen Merkmalen die City-Block-Metrik und bei ordinalen Merkmalen die Rangdifferenz herangezogen. Bei den

nominalen Merkmalen wird zwischen Übereinstimmung und Nicht-Übereinstimmung der Ausprägungen unterschieden.

Neben diesen eben genannten Einflussfaktoren muss für die Studie noch festgelegt werden, in welchem Rahmen das Donor-Limit variiert wird. Neben den beiden Extremfällen, dass einem Spender maximal einmal erlaubt wird zu spenden oder dass er beliebig oft verwendet werden kann, sollen noch zwei weitere Zwischenstufen betrachtet werden, so dass insgesamt die folgenden vier Fälle unterschieden werden:

- Ein Spender wird nur einmal zur Imputation zugelassen.
- Ein Spender wird maximal in 25% der Fälle, in denen er eingesetzt werden könnte, zur Imputation zugelassen.
- Ein Spender wird maximal in 50% der Fälle, in denen er eingesetzt werden könnte, zur Imputation zugelassen.
- Ein Spender kann beliebig oft zur Imputation verwendet werden.

4.1.1.2 Gütekriterien

Zur Beurteilung der Güte der durchgeführten Imputationen werden grundsätzlich die relativen Abweichungen von Verteilungsparametern der imputierten Datenmatrizen zu den entsprechenden Parametern der wahren Datenmatrizen berechnet. Abhängig von der Skalierung der Merkmale werden dabei folgende Verteilungsparameter verwendet (vgl. Nordholt, 1998, S. 163 ff.):

- **Dichotomes Merkmal:** Ausprägungshäufigkeit
- **Ordinales Merkmal:** Median, Quartilsabstand
- **Kardinales Merkmal:** Mittelwert, Varianz

Zur Untersuchung der Unterschiede der vier Fälle einer Begrenzung der Spender im Hinblick auf die Güte der resultierten Imputationen werden dann die Abweichungen der jeweiligen Verteilungsparameter analysiert. Für jeden der vier Fälle wird für alle betrachteten Verteilungsparameter dazu zunächst die relative Abweichung Δp zwischen dem jeweils wahren Verteilungsparameter

p_T und dem entsprechenden, auf Basis der imputierten Daten ermittelten, Verteilungsparameter p_I gemäß

$$\Delta p = \frac{p_I - p_T}{p_T} \tag{4.1}$$

berechnet. Ein negativer Wert für Δp deutet dabei eine Unterschätzung, ein positiver Wert eine entsprechende Überschätzung des jeweiligen wahren Parameters an.

Um die Auswirkungen der unterschiedlichen Fälle einer Begrenzung der Spender auf die Güte der Imputationen zu untersuchen, kann nun ein Vergleich der jeweils resultierenden Werte Δp durchgeführt werden. Aufgrund der in der Simulationsstudie zu erwartenden großen Datenmengen werden jedoch Signifikanztests nicht zielführend sein, so dass alternativ dazu Cohens Effektstärke herangezogen wird (vgl. Cohen, 1997, S. 157, Borz und Döring, 2009, S. 606). Diese ergibt sich gemäß

$$d = \frac{|\Delta\bar{p}_1| - |\Delta\bar{p}_2|}{\sqrt{\frac{s_1^2+s_2^2}{2}}}, \tag{4.2}$$

wobei $\Delta\bar{p}_1$ und $\Delta\bar{p}_2$ die Mittelwerte aller nach Formel (4.1) berechneten relativen Abweichungen für die zwei zu vergleichenden Fälle einer Begrenzung des Spender bezeichnen und s_1^2 beziehungsweise s_2^2 die zugehörigen Varianzen darstellen. Aufgrund der gleichen Stichprobenumfänge ist es nicht notwendig, eine gepoolte Standardabweichung heranzuziehen. Durch die Verwendung der absoluten Werte für $\Delta\bar{p}_1$ und $\Delta\bar{p}_2$ in (4.2) kann auch das Vorzeichen von d interpretiert werden. Ein positives Vorzeichen bedeutet, dass der zweite Fall einer Begrenzung des Spenders besser ist als der erste Fall, während dies bei einem negativen Vorzeichen gerade umgekehrt ist. Gemäß Cohen (1997, S. 157) sind absolute Effektstärken um 0,2 bedeutsam, Fröhlich und Pieter (2009, S. 141) stufen bereits Werte ab 0,1 in diese Kategorie ein.

4.1.1.3 Durchführung der Studie

Für jede Kombination der festgelegten Faktoren „Anzahl der Imputationsklassen“, „Objektanzahl je Imputationsklasse“, „Klassenstruktur“ und „Anzahl der Ausprägungen der ordinalen Merkmale“ werden jeweils 100 vollständige Datenmatrizen erzeugt und die entsprechenden wahren Verteilungsparameter

berechnet. Anschließend werden in jeder dieser 1.600 Datenmatrizen fehlende Daten generiert, wobei jede Kombination der Faktoren „Anteil fehlender Daten“ und „Ausfallmechanismus“ jeweils zehnmal herangezogen wird. Auf Basis dieser 144.000 unvollständigen Datenmatrizen erfolgt dann für jede der vier festgelegten Begrenzungen der Spender eine Imputation mit Hilfe der sechs festgelegten Hot-Deck-Varianten, so dass letztendlich im Rahmen der Simulationsstudie 3.456.000 imputierte Datenmatrizen erzeugt werden. Für diese Datenmatrizen werden abschließend noch die entsprechenden Verteilungsparameter berechnet, um einen Vergleich mit den wahren Werten durchführen zu können.

Bei der Generierung der fehlenden Daten werden die betrachteten Ausfallmechanismen im Rahmen der Studie wie folgt simuliert: Bei MCAR werden die zu löschenden Daten durch zufälliges Ziehen ohne Zurücklegen festgelegt. Demgegenüber wird der Fall MAR dadurch erzeugt, dass abhängig von der Ausprägung des ersten dichotomen Merkmals, bei dem dann keine fehlenden Daten generiert werden, der Anteil fehlender Daten bei den anderen Merkmalen um 10% erhöht beziehungsweise reduziert wird. Um schließlich den Fall NMAR zu simulieren, wird die bei MAR gewählte Vorgehensweise in der Form modifiziert, dass jetzt auch beim ersten dichotomen Merkmal fehlende Daten entsprechend erzeugt werden.

Des Weiteren müssen bei der Generierung der fehlenden Daten noch Einschränkungen vorgenommen werden, um Komplikationen bei der anschließenden Imputation vorzubeugen. Zunächst muss ausgeschlossen werden, dass innerhalb einer Imputationsklasse weniger als 50% der Objekte fehlende Daten aufweisen. Diese Einschränkung ist notwendig, um auch den Fall problemlos abbilden zu können, dass ein Spender nur einmal zur Imputation verwendet wird. Des Weiteren wird der Fall ausgeschlossen, dass bei einem Objekt gleichzeitig die Werte bei allen betrachteten Merkmalen fehlen.

4.1.2 Ergebnisse

Basierend auf den Resultaten der durchgeführten Simulation sollen nun die aufgezeigten Forschungsfragen beantwortet werden. Im ersten Abschnitt wird dazu zunächst untersucht, ob ein Donor-Limit grundsätzlich sinnvoll ist und welche Auswirkungen dieses hat. Der nachfolgende Abschnitt beschäftigt sich

dann mit einer Analyse der Einflussfaktoren, die für oder gegen ein Donor-Limit sprechen. Abschließend wird untersucht, inwieweit Empfehlungen hinsichtlich eines konkreten Wertes eines Donor-Limits ausgesprochen werden können.

4.1.2.1 Auswirkungen des Donor-Limits

Um die Einflussfaktoren hinsichtlich der Häufigkeit einer Verwendung der Spender grundsätzlich zu analysieren, werden zunächst die Effektstärken nach Cohen hinsichtlich des Falls, dass ein Objekt maximal einmal als Spender herangezogen werden kann, und des Falls einer unbeschränkt möglichen Verwendung betrachtet. Für diese Effektstärken bezüglich der einzelnen, in Abschnitt 4.1.1.1 festgelegten Verteilungsparameter, werden dann jeweils der Median sowie einige Streuungsparameter berechnet, die der Tabelle 4.1 entnommen werden können.

	Kardinale Merkmale		Ordinale Merkmale		Nominale Merkmale
	Mittelwert	Varianz	Median	Quartilsabstand	Ausprägungshäufigkeit
Median	-0,001	-0,009	-0,002	-0,008	-0,007
Spannweite	0,104	2,468	0,171	2,231	3,333
Abstand 90%/10%-Quantil	0,031	0,336	0,037	0,262	0,251
Quartilsabstand	0,013	0,068	0,016	0,050	0,045
Standardabweichung	0,013	0,280	0,017	0,249	0,325

Tabelle 4.1: Median und Variabilität der Effektstärken

Bei der Betrachtung dieser Werte fällt zunächst auf, dass keine durchweg positiven oder negativen Effektstärken vorliegen, sondern diese um den Nullpunkt in beide Richtungen streuen. Des Weiteren ist ersichtlich, dass alle Streuungsparameter bei den Verteilungsparametern Mittelwert und Median relativ klein sind, so dass bezüglich dieser Verteilungsparameter nicht mit bedeutenden Effekten zu rechnen ist. Im Vergleich dazu sind bei den Verteilungsparametern Varianz, Quartilsabstand und Ausprägungshäufigkeit sehr hohe Spannweiten für die Effektstärken zu beobachten. Auch der Abstand zwischen 90%- und 10%-Quantil sowie die Standardabweichung sind bei diesen Verteilungsparametern relativ hoch. Allerdings wird bei Betrachtung des Quartilsabstands in Kombination mit dem Median auch deutlich, dass ein Großteil der Effekte als

trivial eingestuft werden muss. Daraus ist insgesamt ersichtlich, dass die Verwendungshäufigkeit der Spender grundsätzlich zwar bedeutsame Einflüsse auf die Güte der Imputation hat, dieser Effekt aber nicht durchgängig auftritt. Somit wird es von Interesse sein zu untersuchen, in welchen Situationen bedeutsame Effekte vorliegen, worauf im nächsten Abschnitt noch eingegangen wird.

Der durch kombinatorische Überlegungen resultierende Stand der Forschung, dass im Fall einer maximal einmaligen Verwendung eines Spenders die Varianz der geschätzten Verteilungsparameter der vervollständigten Daten reduziert wird, soll an dieser Stelle auf Basis der Simulationsergebnisse noch empirisch untersucht werden. In der Tabelle 4.2 sind dazu die prozentualen Häufigkeiten angegeben, in wie vielen Fällen einer entsprechenden Parameterschätzung innerhalb einer Faktorstufenkombination eine der vier Donor-Limits zur kleinsten Varianz der Schätzwerte führt.

		Donor-Limit			
Betrachteter Parameter		einmal	25%	50%	unbegrenzt
Kardinale Merkmale	Mittelwert	68,52%	15,47%	7,95%	8,06%
	Varianz	67,25%	15,74%	8,56%	8,45%
Ordinale Merkmale	Median	74,54%	11,38%	7,62%	6,46%
	Quartilsabstand	85,88%	5,71%	4,96%	3,45%
Dichotome Merkmale	Ausprägungshäufigkeit	78,36%	8,41%	6,96%	6,27%

Tabelle 4.2: Häufigkeitsverteilung der minimalen Varianz der Schätzwerte

Es ist deutlich zu erkennen, dass in den meisten Fällen eine maximal einmalige Verwendung eines Spenders zur kleinsten Variabilität der geschätzten Parameter führt und es damit zu einer Verbesserung der Schätzgenauigkeit kommt. Dieser Sachverhalt kommt bei den ordinalen und dichotomen Merkmalen noch stärker zum Ausdruck als bei den kardinalen Merkmalen. Allerdings ist auch festzuhalten, dass eine häufigere beziehungsweise nicht eingeschränkte Verwendung der Spender in einer Reihe von Faktorstufenkombinationen durchaus eine minimale Variabilität des geschätzten Parameters zur Folge hat.

4.1.2.2 Analyse der Einflüsse auf die Vorteilhaftigkeit des Donor-Limits

Um die Einflussfaktoren hinsichtlich der Häufigkeit einer Verwendung der Spender grundsätzlich zu analysieren, werden wiederum die Effektstärken nach Cohen (1997) zwischen den Fällen Donor-Limit von Eins und unbegrenzt betrachtet. Negative Werte sprechen somit für ein Donor-Limit, während positive Werte andeuten, dass kein Donor-Limit von Vorteil ist. Bei der nachfolgenden Darstellung der Ergebnisse werden zunächst die Haupteffekte und darauf aufbauend vorliegende Wechselwirkungen der Einflussfaktoren untersucht.

In der Tabelle 4.3 sind die Effektstärken für die betrachteten Verteilungsparameter in Abhängigkeit der in dieser Studie betrachteten Einflussfaktoren zusammengefasst. Effektstärken, die im Betrag einen Wert ab 0,1 aufweisen, sind fett hervorgehoben.

Bei einer Betrachtung der Ergebnisse fällt zunächst auf, dass bei den Lageparametern für kardinale und ordinale Merkmale auch unabhängig von den betrachteten Einflussgrößen nur triviale Effekte vorliegen, das heißt, ein Donor-Limit hat für keine Ausprägung der Faktoren einen bedeutsamen Einfluss auf die Schätzgenauigkeit dieser Verteilungsparameter. Dieses Ergebnis deckt sich auch mit den Erkenntnissen aus dem vorherigen Abschnitt.

Demgegenüber können bei den Streuungsparametern sowie der Ausprägungshäufigkeit bei einigen Faktoren bedeutsame Effekte festgestellt werden. Dabei fällt auf, dass bei einem hohen Prozentsatz fehlender Daten sowie der Hot-Deck-Variante „SimDL“ ein Donor-Limit durchgängig zu besseren Schätzergebnissen führt. Auch eine höhere Anzahl von Imputationsklassen spricht tendenziell für ein Donor-Limit.

Während die Effekte der Dimension der Datenmatrix sowie der Objektanzahl je Imputationsklasse nicht eindeutig sind, spielt es hinsichtlich der vorliegenden Klassenstruktur sowie bei Verwendung der zufälligen Hot-Deck-Varianten und der Variante „SeqD“ keine Rolle, ob eine Beschränkung der Spenderverwendungshäufigkeit erfolgt oder nicht. Auffällig ist noch die Tatsache, dass die Hot-Deck-Variante „SimD“ teilweise positive Effektgrößen aufweist, die für eine unbegrenzte Verwendungsmöglichkeit der Spender sprechen. Hier könnte die Analyse von Wechselwirkungseffekten zu weiteren interessanten Erkenntnissen führen.

Faktor	Faktorstufen	Kardinale Merkmale: Mittelwert	Kardinale Merkmale: Varianz	Ordinale Merkmale: Median	Ordinale Merkmale: Quartilsabstand	Nominale Merkmale: Ausprägungshäufigkeit
Dimension der Datenmatrix	(100×9)	0,000	-0,082	-0,001	-0,030	-0,034
	(350×9)	0,000	**-0,177**	-0,005	**-0,152**	-0,022
	(500×9)	0,000	-0,064	-0,004	-0,030	**-0,130**
	(1750×9)	0,001	**-0,146**	-0,006	-0,065	**-0,162**
Anzahl der Imputationsklassen	2	0,000	-0,068	-0,001	-0,029	-0,072
	7	0,000	**-0,147**	-0,003	**-0,115**	-0,090
Objekte je Imputationsklasse	50	0,000	**-0,112**	-0,001	-0,073	-0,028
	250	0,000	-0,090	-0,005	-0,041	**-0,141**
Klassenstruktur	Stark	0,000	-0,092	-0,001	-0,072	-0,072
	Schwach	0,000	-0,094	-0,001	-0,045	-0,080
Prozentsatz fehlender Daten	5%	0,000	-0,025	0,000	-0,013	-0,011
	10%	0,000	-0,071	0,000	-0,037	-0,051
	20%	0,000	**-0,148**	0,000	**-0,100**	**-0,129**
Ausfallmechanismus	MCAR	0,001	-0,088	-0,001	-0,053	-0,065
	MAR	0,000	**-0,100**	0,000	-0,066	-0,086
	NMAR	0,001	-0,091	0,000	-0,058	-0,077
Hot-Deck-Verfahren	SimD	-0,001	**0,153**	-0,002	0,025	0,075
	SimDL	-0,004	**-0,339**	0,005	**-0,214**	**-0,338**
	SeqD	0,001	-0,007	-0,003	0,000	-0,005
	SeqDL	0,000	-0,088	0,010	**-0,133**	-0,041
	SimZ	0,000	-0,001	-0,001	-0,004	0,000
	SeqZ	0,000	-0,001	0,000	-0,001	-0,003

Tabelle 4.3: Effektstärken in Abhängigkeit der Einflussfaktoren

Auf Basis der aus der Analyse der Haupteffekte gewonnenen Erkenntnisse sollen nun zunächst die Effektstärken für die Verteilungsparameter in Abhängigkeit aller Kombinationen zwischen den Hot-Deck-Varianten „SimD“, „SimDL“ und „SeqDL“ sowie der restlichen Einflussfaktoren untersucht werden. Die entsprechenden Werte sind dazu in der Tabelle 4.4 dargestellt, wobei Effektstärken ab 0,1 wiederum hervorgehoben sind.

Wie bei der Analyse der Haupteffekte zeigt sich auch jetzt, dass beim Verfahren „SimD“ ein Donor-Limit nicht von Vorteil ist. Über alle Kombinationen mit den anderen Faktoren ergeben sich mit einer Ausnahme positive Werte, wenngleich nur bei der Varianz und der Ausprägungshäufigkeit bedeutsame Effekte vorliegen. Des Weiteren ist zu erkennen, dass bei den beiden Verfahren „SimDL“ und „SeqDL“ ausnahmslos negative Werte auftreten, die darüber

hinaus größtenteils auf bedeutsame Effekte hinweisen und damit für das restriktivste Donor-Limit sprechen.

		SimD			SimDL			SeqDL		
Faktor	Faktorstufen	V	Q	A	V	Q	A	V	Q	A
Dimension der Datenmatrix	(100 × 9)	**0,140**	0,053	0,058	**-0,337**	**-0,192**	**-0,216**	-0,089	**-0,139**	-0,026
	(350 × 9)	**0,235**	0,058	0,055	**-0,473**	**-0,333**	**-0,278**	**-0,120**	**-0,207**	-0,018
	(500 × 9)	**0,120**	0,040	**0,111**	**-0,283**	**-0,116**	**-0,492**	-0,077	-0,064	-0,073
	(1750 × 9)	**0,215**	0,045	**0,108**	**-0,420**	**-0,257**	**-0,554**	**-0,109**	**-0,132**	-0,064
Anzahl der Imputationsklassen	2	0,097	0,025	0,081	**-0,247**	**-0,101**	**-0,300**	-0,066	-0,082	-0,049
	7	**0,287**	0,033	0,075	**-0,521**	**-0,382**	**-0,424**	**-0,130**	**-0,217**	-0,031
Objekte je Imputationsklasse	50	**0,182**	0,082	0,034	**-0,426**	**-0,284**	**-0,132**	**-0,111**	**-0,196**	-0,004
	250	**0,143**	0,056	**0,140**	**-0,319**	**-0,131**	**-0,684**	-0,088	-0,047	-0,098
Klassen-struktur	stark	**0,153**	0,048	0,078	**-0,339**	**-0,156**	**-0,362**	-0,091	**-0,135**	-0,042
	schwach	**0,144**	0,006	0,071	**-0,338**	**-0,269**	**-0,313**	-0,085	**-0,132**	-0,040
Prozentsatz fehlender Daten	5%	0,065	-0,012	0,031	-0,084	-0,057	-0,045	-0,013	-0,028	-0,004
	10%	**0,148**	0,006	0,077	**-0,262**	**-0,162**	**-0,213**	-0,039	-0,073	-0,010
	20%	**0,203**	0,061	**0,101**	**-0,558**	**-0,345**	**-0,600**	**-0,168**	**-0,233**	-0,085
Ausfallmechanismus	MAR	**0,151**	0,025	0,079	**-0,355**	**-0,226**	**-0,372**	**-0,107**	**-0,152**	-0,058
	MCAR	**0,153**	0,023	0,067	**-0,326**	**-0,204**	**-0,296**	-0,075	**-0,119**	-0,025
	NMAR	**0,154**	0,029	0,077	**-0,334**	**-0,213**	**-0,344**	-0,081	**-0,125**	-0,038

Tabelle 4.4: Wechselwirkungen zwischen Imputationsmethode und restlichen Faktoren (Legende: V = Varianz, Q = Quartilsabstand, A = Ausprägungshäufigkeit)

Für alle drei betrachteten Hot-Deck-Varianten ist ersichtlich, dass bei einer hohen Anzahl von Imputationsklassen sowie einem höheren Prozentsatz fehlender Daten bedeutsame Effekte vorliegen. Bezüglich der Anzahl der Objekte je Imputationsklasse zeigt sich demgegenüber kein einheitlicher Effekt, da je nach Hot-Deck-Variante und Skalierung der Merkmale eine geringere wie auch eine höhere Anzahl an Objekten je Klasse einen bedeutenden Effekt hervorruft. Bei den restlichen Einflussfaktoren lassen sich unabhängig von den jeweiligen Ausprägungen bedeutsame und triviale Effekte gleichermaßen feststellen, die damit nur auf das Imputationsverfahren beziehungsweise die Skalierung der Merkmale zurückzuführen sind.

Neben den nach den Hot-Deck-Varianten differenziert betrachteten Einflüssen liegen auch einige Wechselwirkungen höherer Ordnung vor, die zu auffällig großen absoluten Werten für die Effektstärke führen. Beispielsweise treten im Fall 20% fehlender Werte sowie einer hohen Anzahl von Imputationsklassen und einer geringen Objektanzahl in diesen Klassen bei dem Verfahren „SimDL“ Effektstärken bis zu $-1{,}7$ für die Varianz und bis zu $-1{,}9$ beim Quartilsabstand auf. Effektstärken bis zu einem Wert von -3 ergeben sich für die Ausprägungs-

häufigkeit im Fall einer hohen Klassenanzahl mit vielen Objekten je Klasse und einem hohen Anteil fehlender Daten. Betrachtet man demgegenüber das Verfahren „SimD," so sind die stärksten Effekte im Fall einer hohen Klassenanzahl mit wenigen Objekten je Klasse und einem hohen Anteil fehlender Daten mit Werten von maximal 0,6 bei der Varianz und 0,34 beim Quartilsabstand deutlich kleiner. Auffällig ist dennoch, dass insbesondere die Kombination aus Hot-Deck-Variante, Anzahl der Imputationsklassen, Objekte je Klasse und Anteil fehlender Daten sehr bedeutsame Effektstärken zur Folge hat, die zum Teil für das restriktivste Donor-Limit als auch kein Donor-Limit spricht. Darüber hinaus bestätigt die Analyse der Wechselwirkungen höherer Ordnung die für einzelne Hot-Deck-Verfahren bereits festgestellte, jeweils vorteilhafte, Form des Donor-Limits.

4.1.2.3 Analyse der Donor-Limits

Bislang wurde im Wesentlichen untersucht, ob und unter welchen Bedingungen kein Donor-Limit oder die restriktivste Form des Donor-Limits zu besseren Ergebnissen führt. Jetzt sollen auch Donor-Limits zwischen diesen beiden Extremfällen in die Betrachtung aufgenommen werden. Dazu wird ermittelt, wie häufig einer der in dieser Studie betrachteten vier Fälle des Donor-Limits die beste Parameterschätzung liefert. Die entsprechenden prozentualen Häufigkeiten sind in der Tabelle 4.5 dargestellt.

Betrachteter Parameter		Donor-Limit einmal	25%	50%	unbegrenzt
Kardinales Merkmal	Mittelwert	42,71%	20,22%	18,48%	18,60%
	Varianz	54,05%	17,79%	13,04%	15,12%
Ordinales Merkmal	Median	46,41%	21,53%	14,47%	17,59%
	Quartilsabstand	56,83%	16,24%	12,94%	13,99%
Dichotomes Merkmal	Ausprägungshäufigkeit	49,42%	18,94%	15,07%	16,57%

Tabelle 4.5: Häufigkeitsverteilung der geringsten Abweichung vom wahren Verteilungsparameter

Es zeigt sich, dass in den meisten Fällen ein Donor-Limit von Eins zur besten Parameterschätzung führt. Dieser Sachverhalt kommt bei den Variabilitätsmaßen stärker zum Ausdruck als bei den Lageparametern. Dabei ist durchgängig zu erkennen, dass die Häufigkeiten über die vier betrachteten Fälle eines

Donor-Limits zunächst ab- und danach wieder zunehmen, so dass nochmals deutlich wird, dass situationsbedingt unterschiedliche Beschränkungen jeweils von Vorteil sind.

Auf Basis dieser Ergebnisse kann festgehalten werden, dass grundsätzlich ein Optimierungspotenzial bezüglich der konkreten Anzahl, wie häufig ein Spender maximal herangezogen werden soll, erkennbar ist. An dieser Stelle ist es nicht sinnvoll, die zugrundeliegenden Zusammenhänge näher zu analysieren und konkrete Empfehlungen diesbezüglich abzuleiten. Die hier gewonnenen Erkenntnisse sprechen jedoch eindeutig dafür, entsprechende zukünftige Forschungsbemühungen auf diesen Bereich zu fokussieren.

4.1.3 Zusammenfassung

Die im Rahmen dieses Abschnitts durchgeführte Simulationsstudie zeigt, dass es deutliche Unterschiede zwischen Hot-Deck-Imputationen gibt, bei denen das Donor-Limit variiert wird. Ein Donor-Limit ist nicht grundsätzlich von Vorteil, da situationsbedingt auch eine unbeschränkte Spenderverwendung zu besseren Ergebnissen führen kann.

Einige Gegebenheiten des vorliegenden Datenmaterials sprechen im Hinblick auf dadurch verbesserte Parameterschätzungen für ein Donor-Limit. Falls die Anzahl der Imputationsklassen gering ist, ergibt sich ein Vorteil hinsichtlich der Schätzgenauigkeit der Streuungsparameter bei kardinalen und ordinalen Merkmalen. Bei eher wenigen Objekten je Imputationsklasse profitiert die Varianz der kardinalen Merkmale von einem Donor-Limit, während bei den dichotomen Merkmalen viele Objekte für ein Donor-Limit sprechen. Darüber hinaus ist dies grundsätzlich bei einer hohen Anzahl von fehlenden Daten der Fall. Insgesamt kann auch festgehalten werden, dass die Schätzung der Lageparameter der kardinalen und ordinalen Merkmale durch ein Donor-Limit nicht nennenswert beeinflusst wird.

Neben den Gegebenheiten des Datenmaterials stellt vor allem die verwendete Hot-Deck-Variante einen bedeutenden Einflussfaktor dar. Je nach Verfahrensvariante bringt ein Donor-Limit Vorteile oder Nachteile mit sich beziehungsweise bleibt gegebenenfalls auch ohne nennenswerten Einfluss auf die Parameterschätzungen. Bei den beiden zufälligen Imputationsverfahren sowie der Variante „SeqD“ zeigt eine Begrenzung nie bedeutsame Effekte. Demgegenüber

ist im Fall der Verfahrensvarianten „SeqDL“ und „SimDL“ eine Begrenzung sinnvoll, während bei „SimD“ kein Donor-Limit zu empfehlen ist.

Auch wenn in den meisten Fällen eine maximal einmalige Verwendung eines Spenderobjekts zur besten Parameterschätzung führt, sind situationsbedingt auch keine Beschränkungen von Vorteil. Die Bestimmung, in welchen Fällen konkret ein Donor-Limit von Vorteil ist, erscheint somit sinnvoll, so dass detaillierte Untersuchungen der diesbezüglich zugrundeliegenden Zusammenhänge einen sehr interessanten Ansatz für zukünftige Forschungsarbeiten darstellen könnten.

4.2 Konfirmatorische Studie

Im Hinblick auf eine abschließende Betrachtung der Auswirkungen eines Donor-Limits auf die Ergebnisse einer Imputation wird im Folgenden eine konfirmatorische Simulationsstudie durchgeführt. Ziele dieser Studie sind eine tiefgründige Analyse einzelner Aspekte der Voruntersuchung sowie eine Überprüfung der Belastbarkeit jener Aussagen, die basierend auf der Voruntersuchung gemacht wurden. Des Weiteren soll der Einfluss weiterer Faktoren, welche das Datenmaterial charakterisieren können, auf die Vorteilhaftigkeit eines Donor-Limits untersucht werden.

Diesen Fragestellungen wird im Rahmen der nächsten Abschnitte nachgegangen. Hierzu wird in Abschnitt 4.2.1 das Studiendesign vorgestellt. In dem darauf folgenden Abschnitt 4.2.2 werden die Ergebnisse der Simulation präsentiert. Im letzten Abschnitt werden die Ergebnisse zusammengefasst und allgemeine Aussagen über Konstellationen, unter denen ein Donor-Limit für die Imputationsqualität von Vorteil ist, getroffen.

4.2.1 Studiendesign

Im Rahmen dieser Simulationsstudie soll untersucht werden, ob die Empfehlungen und Aussagen der Voruntersuchung auch unter leicht veränderten Bedingungen Bestand haben. Zudem sollen die betrachteten Konstellationen tiefgehender analysiert werden. Hierzu erfolgt im nächsten Abschnitt eine Darstellung jener Faktoren, die in dieser Simulationsstudie variiert werden. Daraufhin werden bedeutende Eckpunkte der Implementierung dieser Simulationsstudie,

welche unter anderem Auswirkungen auf die Berechnung der Gütekriterien haben, vorgestellt. Abgeschlossen wird dieser Abschnitt mit einer Darstellung der verwendeten Gütekriterien, mittels derer die gestellten Forschungsfragen beantwortet werden.

4.2.1.1 Einflussfaktoren

Die Einflussfaktoren, welche für diese konfirmatorische Studie gewählt wurden, orientieren sich grundsätzlich an jenen der Voruntersuchung und somit an denen von Roth und Switzer III (1995), Roth et al. (1999) und Strike et al. (2001). Eine vollständige Replizierung der Voruntersuchung erscheint bei Simulationsstudien jedoch nicht geeignet, da keine namhaften Veränderungen in den Ergebnissen zu erwarten sind. Vielmehr werden die Stufen der grundlegenden Faktoren gezielt mittels der Erkenntnisse aus der Voruntersuchung verändert, um aussagekräftige und stabile Ergebnisse zu erzeugen. Im Einzelnen stellen sich die verwendeten Faktoren und die Veränderungen der Faktorstufen gegenüber der Voruntersuchung wie folgt dar:

Anzahl der Objekte: Die Anzahl der vorliegenden Objekte ist für die wirksame Imputation mittels eines Hot-Deck-Verfahrens zentral. Je mehr Objekte vorhanden sind, desto wahrscheinlicher ist es, dass mindestens ein Spender existiert, der dem Empfänger hinreichend ähnlich ist. Des Weiteren kann erwartet werden, dass bei mehr Objekten jener Nachteil, der sich durch die Wahl eines zweit- oder dritt-besten Spenders ergibt, abnimmt[1]. Folglich kann vermutet werden, dass mit steigender Objektanzahl die Nachteile des Donor-Limits geringer werden.

Bei Festlegung der Objektanzahl muss beachtet werden, dass im Zusammenspiel mit den betrachteten Anteilen an fehlenden Werten eine Ganzzahligkeit gefordert ist. Um dieser nachzukommen, wird die Anzahl an Objekten auf 100 beziehungsweise 200 festgelegt.

Merkmalsskalierung: Es werden gemischt skalierte Datenmatrizen betrachtet. Die Matrizen bestehen jeweils aus der gleichen Anzahl an nominalen, ordinalen und quantitativen Kovariaten. Als nominal skalierte Merkmale werden ausnahmslos dichotome Merkmale betrachtet, da

[1] Sofern alle anderen Bedingungen, insbesondere der Anteil fehlender Werte, gleich bleiben.

jedes nominal polytom skalierte Merkmal durch eine Anzahl an dichotomen Merkmalen codiert werden kann. Die Darstellung der ordinalen Merkmale erfolgt mittels der Ausprägungen eins bis sieben, ähnlich der in den Sozialwissenschaften häufig verwendeten sieben Punkte Likert-Skala. Quantitative Merkmale sind mit einem Erwartungswert von Null und einer Standardabweichung von Eins normal verteilt.

Anzahl der Merkmale: Die Anzahl der Kovariaten, die zur Berechnung der Spender-Empfänger-Distanzen dienen, wird auf Drei oder Neun festgelegt. Im Folgenden werden vollständige Kovariaten simuliert, um von dem Effekt der verschiedenen Distanzberechnungsmethoden zu abstrahieren. Denn im Idealfall sollten Methoden zur Distanzberechnung unter Nutzung unvollständiger Merkmale zu denselben Distanzen führen wie bei Vollständigkeit der Kovariaten.

Die jeweilige Anzahl an Merkmalen mit fehlenden Werten wird auf Eins festgelegt, dessen Skalierung variiert wird. Unterschieden wird hierbei, wie bei den Kovariaten, zwischen nominal, ordinal und quantitativ.

Eine Erhöhung der Merkmalsanzahl stellt eine Verbesserung der vorhandenen Informationen zur Berechnung der Spender-Empfänger-Distanzen dar. Diese Variation erfolgt, da mit zunehmendem Informationsniveau zwei Entwicklungen erwartet werden können. Zum einen führen mehr vorhandene Informationen im Allgemeinen zu einer verbesserten Spenderidentifizierung. Zum anderen wird die Wahrscheinlichkeit, dass ein Spender als nächster Nachbar zu zwei oder mehr Empfängern bewertet wird, reduziert. Diese beiden Entwicklungen würden dazu führen, dass mit zunehmendem Informationsniveau die Imputationsqualität steigt und sich einem Optimum nähert.

Anzahl der Imputationsklassen: Aus den Ergebnissen der Voruntersuchung geht hervor, dass durch eine Erhöhung der Imputationsklassenanzahl der komparative Vorteil der überlegenen Vorgehensweise steigt (vgl. Tabellen 4.3 und 4.4). Grundsätzlich ist jenes Verfahren, das für wenig Imputationsklassen geeigneter ist, auch immer besser, wenn eine

Datenmatrix in mehr Imputationsklassen unterteilt wird[2]. Demzufolge kann erwartet werden, dass ein Hot-Deck-Verfahren, welches innerhalb einer einzelnen Klasse bessere Imputationsergebnisse liefert, auch insgesamt bessere Imputationen vornimmt, da jeweils ein kleinerer Fehler propagiert wird. Folglich scheidet eine Notwendigkeit, die Anzahl der Imputationsklassen zu variieren, aus, da der Effekt, für sinnvolle Möglichkeiten Imputationsklassen zu bilden, eindeutig ist. Somit sollten die generierten Datenmatrizen als einzelne Imputationsklasse betrachtet werden.

Daten-/Klassenstruktur: Wie auch in der Voruntersuchung soll die Struktur der Daten variiert werden. Dieser Faktor wird, anstelle von zwei, auf drei Stufen variiert. Unterschieden wird zwischen einer schwachen, mittleren und starken Struktur. Die Erzeugung dieser Strukturen wird mittels Pearson-Korrelationen von $\rho = 0{,}00$; 0,35 und 0,70 zwischen allen Merkmalen sichergestellt.

Anteil fehlender Werte: Gemäß den Ergebnissen der Voruntersuchung hat der Anteil fehlender Werte einen maßgeblichen Einfluss auf die Vorteilhaftigkeit eines Donor-Limits. Daher wird im Folgenden der Anteil fehlender Werte zum besseren Verständnis der auftretenden Effekte in einem größeren Bereich auf mehr Stufen variiert. Betrachtet werden die Prozentsätze von fehlenden Werten 5% bis 50% in 5% Schritten. Eine Analyse dieser zehn Faktorstufen wird insbesondere Aufschluss über die Imputationsqualität unter Extremsituationen geben.

Ausfallmechanismus: Es wird zwischen drei Ausfallmechanismen in ansteigender Intensität unterschieden. Der erste betrachtete Ausfallmechanismus ist *MCAR*. Hier wird zufällig der vorgegebene Anteil an Werten gelöscht. Für die MAR-Mechanismen wird, basierend auf dem Wert einer dichotomen Kovariate, der Datensatz zweigeteilt. Der erste MAR-Ausfallmechanismus löscht Daten, so dass die Menge an fehlenden Werten in einem Teil immer doppelt so hoch ist wie in dem anderen, jedoch, über die gesamte Datenmatrix betrachtet, der Anteil fehlender Werte

[2] Unter der Voraussetzung, dass die Imputationsklassen sinnvoll gebildet werden (vgl. hierzu Abschnitt 3.2.1.1).

den Vorgaben entspricht. Der zweite MAR-Ausfallmechanismus ist dem ersten ähnlich, jedoch ist der Ausfall in dem zweiten Teil der Daten vier mal so groß wie in dem ersten. Konzeptionell entsprechen diese Ausfallmechanismen jenen in Abschnitt 3.3.2 und werden deshalb auch mit *MAR 1:2* und *MAR 1:4* bezeichnet. Auf die Betrachtung eines NMAR-Ausfallmechanismus wird zugunsten des intensiveren MAR-Mechanismus verzichtet, da im Allgemeinen in der Literatur von Imputation bei Vorliegen eines NMAR-Ausfallmechanismus abgeraten wird.

Hot-Deck-Verfahren: Da in der Voruntersuchung die Ergebnisse der zufälligen Hot-Deck-Verfahren nie von einem gesetzten Donor-Limit beeinflusst wurden, beschränkt sich die konfirmatorische Studie auf die Betrachtung der deterministischen Hot-Deck-Verfahren. Hierbei wird für die Bestimmung der Spender-Empfänger-Ähnlichkeiten auf Distanzen abgestellt. Zur Berechnung der merkmalsweisen Distanzen werden bei den quantitativen Merkmalen die City-Block-Metrik und bei den ordinalen Merkmalen die Rangdifferenzen herangezogen. Bei den nominalen Merkmalen wird zwischen Übereinstimmung und Nicht-Übereinstimmung der Ausprägungen unterschieden. Vor der linearen Aggregation der einzelnen Spender-Empfänger-Distanzen werden die Merkmalsdistanzen mit Hilfe des Kehrwertes der entsprechenden maximalen Distanzen gewichtet. Da lediglich ein univariates Ausfallmuster betrachtet wird, ist eine Variation zwischen simultanen und sequentiellen Hot-Deck-Verfahren entbehrlich, da beide in denselben Ergebnissen resultieren.

Donor-Limit: Zur näheren Analyse, ob ein Donor-Limit für eine gegebene Konstellation der variierten Faktoren besser ist, wird zudem zwischen einem Donor-Limit von Eins und einem Donor-Limit von unendlich variiert.

4.2.1.2 Durchführung der Studie

Zur Beantwortung der gestellten Forschungsfragen wird die Imputationsqualität für simulierte Daten verglichen. Die Simulation wurde in der Statistiksoftware R Version 2.15.2 (R Core Team, 2013) nach dem in Abbildung 4.1 dargestellten Schema implementiert.

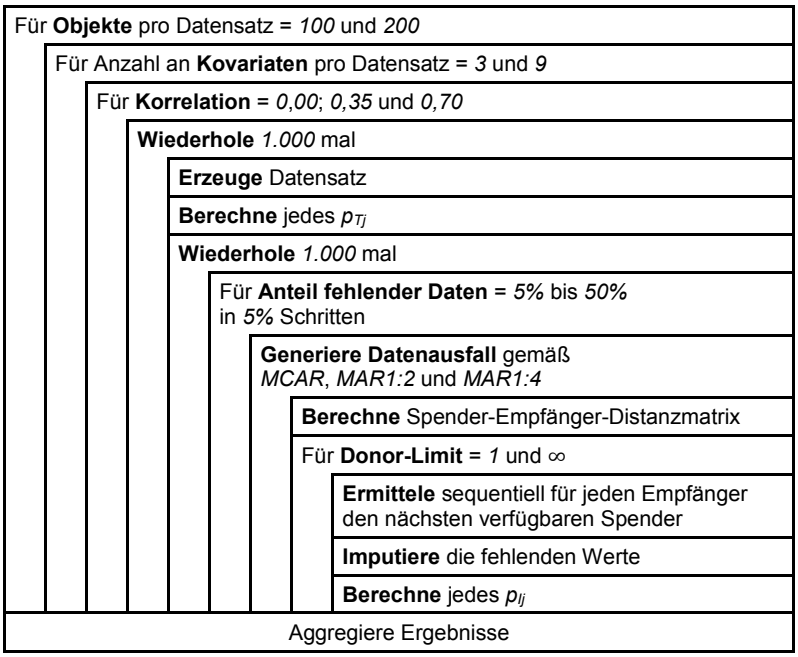


Abbildung 4.1: Umfassende Darstellung des Simulationsdesigns der konfirmatorischen Studie

Grundlage der simulierten Datenmatrizen sind Zufallszahlen, die einer multivariaten Standardnormalverteilung folgen. Zur Generierung dieser (Pseudo-) Zufallszahlen wird die Funktion *rmvnorm* aus dem R-Paket **mvtnorm** (Genz et al., 2013) verwendet. Generiert werden jeweils die gegebene Anzahl an Objekten und ein Merkmal mehr als die gewünschte Anzahl an Kovariaten mit der gegebenen Korrelation[3]. Von den generierten Kovariaten wurde jeweils ein Drittel in nominale und ein Drittel in ordinale Merkmale mittels der NOTRA-Methode transformiert (Cario und Nelson, 1997). Des Weiteren wird jene Variable, bei der Fehlen induziert wird, jeweils nominal-, ordinal- und untransformiert betrachtet. Zur Transformation der nominalen Merkmale werden alle Merkmalsausprägungen kleiner als Null mit 0 und alle größer als Null mit 1 kodiert. Die ordinalen Merkmale werden jeweils durch eine Umkodierung von Merkmalsausprägungen in den Bereichen $(-\infty; -1{,}5]$, $(-1{,}5; -1]$, $(-1; -0{,}5]$, $(-0{,}5; 0{,}5]$, $(0{,}5; 1]$, $(1; 1{,}5]$ und $(1{,}5; \infty)$ zu den Werten 1 bis 7 transformiert. Diese Datenmatrizen werden 1.000 mal generiert.

[3] Zur Einbringung der Korrelationsstruktur in die normalverteilten Vektoren verwendet das R-Paket **mvtnorm** die Eigenwert-Dekompositions-Methode.

Die 12.000 hieraus resultierenden vollständigen Datenmatrizen werden dann den drei beschriebenen Ausfallmechanismen ausgesetzt, jeweils mit 10 unterschiedlichen Anteilen von fehlenden Werten. Dies wird 1.000 mal wiederholt, so dass 360.000.000 Matrizen mit fehlenden Werten entstehen.

Die Distanzen zwischen den Spendern und Empfängern werden mit Hilfe einer Manhattan-Distanz über die Kovariaten berechnet. Zur merkmalsspezifischen Gewichtung werden die Kehrwerte der Merkmalsspannweiten verwendet. Dieses Vorgehen liefert äquivalente Ergebnisse zu der in Abschnitt 4.2.1.1 beschriebenen Methode, und die Imputationsalgorithmen sind über die Funktion *impute.NN_HD*, als Teil des R-Pakets **HotDeckImputation** (Joenssen, 2013), verfügbar. Zuordnungen zwischen den Spendern und Empfängern werden unter Berücksichtigung der zwei betrachteten Donor-Limits durchgeführt. Dies resultiert in 720.000.000 Spender-Empfänger-Zuordnungen, welche jeweils zur Erstellung einer imputierten Datenmatrix verwendet werden. In Summe werden in der Simulation Gütekriterien für 12.000 vollständige und 720.000.000 imputierte Datenmatrizen, oder ca. 47,4 Terabyte an Daten, berechnet.

4.2.1.3 Gütekriterien

Damit die Forschungsfragen beantwortet werden können, müssen diese zunächst für die Simulationsstudie operationalisiert werden. Zur Beantwortung beider Fragen ist grundsätzlich eine Beurteilung der Imputationsgüte notwendig. Diese wird im Folgenden durch die Abweichungen von Verteilungsparametern der imputierten Datenmatrizen zu den entsprechenden Parametern der wahren Datenmatrizen bestimmt.

Die Auswahl an Parametern orientiert sich an der Voruntersuchung. Da festgestellt wurde, dass weder die Schätzung des Mittelwerts eines kardinal skalierten Merkmals noch der Median eines ordinal skalierten Merkmals nennenswert durch eine Veränderung des Donor-Limits beeinflusst wird, wird die Schätzung dieser Parameter im Folgenden nicht weiter betrachtet. Es erfolgt eine Fokussierung auf jene Parameter, mittels derer die Variabilität und Zusammenhänge in den Daten gemessen werden. Abhängig von der Skalierung des zu imputierenden Merkmals werden folgende univariate Verteilungsparameter verwendet:

- **Kardinales Merkmal:** Varianz
- **Ordinales Merkmal:** Quartilsabstand
- **Dichotomes Merkmal:** Ausprägungshäufigkeit

Des Weiteren werden, abhängig von der Merkmalsskalierung, folgende Zusammenhangsmaße verwendet:

- **Kardinales Merkmal:** Korrelationskoeffizient nach Pearson zu einem anderen kardinalen Merkmal
- **Ordinales Merkmal:** Rangkorrelationskoeffizient von Spearman zu einem anderen ordinalen Merkmal
- **Dichotomes Merkmal:** normierter Kontingenzkoeffizient zu einem anderen nominalen Merkmal.

Für die beiden Donor-Limits werden zunächst in einem einzelnen Iterationsschritt j der Simulation folgende Größen berechnet:

$$|\Delta p_j^1| = |p_{Tj} - p_{Ij}^1|$$
$$|\Delta p_j^\infty| = |p_{Tj} - p_{Ij}^\infty|,$$

wobei p_{Tj} die jeweils wahren Verteilungsparameter und p_{Ij}^1 beziehungsweise p_{Ij}^∞ die auf Basis der imputierten Daten ermittelten Verteilungsparameter darstellen. Die Superskripte 1 und ∞ entsprechen dem jeweils verwendeten Donor-Limit.

Um die Auswirkungen des Donor-Limits auf die Imputationsgüte zu untersuchen, wird ein Vergleich der jeweils resultierenden Werte $|\Delta p_j^1|$ beziehungsweise $|\Delta p_j^\infty|$ durchgeführt. Aufgrund der in dieser Simulationsstudie großen Datenmenge werden selbst kleinste Effekte statistisch signifikant sein. Daher stellen Signifikanztests an dieser Stelle keine sinnvolle Methode zur Auswertung dar. Vielmehr wird wiederum auf Cohens d und die hiermit assoziierten Schwellenwerte zurückgegriffen (vgl. Cohen, 1997, S. 157; Borz und Döring, 2009, S. 606).

Abgesehen von dem Einfluss der Stichprobengröße ist Cohens d (vgl. Formel (4.2)) identisch mit der t-Statistik eines Zweistichproben-t-Tests für unabhängige Stichproben. Anhand der Ausführungen in Abschnitt 4.2.1.2, insbesondere

Abbildung 4.1, wird deutlich, dass die $|\Delta p_j^1|$ und $|\Delta p_j^\infty|$ paarweise verbunden sind. Für verbundene Stichproben stellt jedoch der Zweistichproben-t-Test für abhängige Stichproben die mächtigere Alternative dar, welche besser dazu geeignet ist Unterschiede festzustellen. Daher wird im Folgenden eine abgewandelte Effektgröße auf die paarweisen Differenzen $\Delta p'_j = |\Delta p_j^1| - |\Delta p_j^\infty|$ wie folgt berechnet:

$$d' = \frac{\Delta\bar{p}'}{\sqrt{s^{2\prime}}}. \tag{4.3}$$

Der Mittelwert aller $\Delta p'_j$ wird mit Hilfe von

$$\Delta\bar{p}' = \frac{1}{1.000.000} \sum_{j=1}^{1.000.000} \Delta p'_j \tag{4.4}$$

bestimmt. Die Varianz der $\Delta p'_j$ wird unter der Nutzung des Satzes von Steiner mittels

$$s^{2\prime} = \frac{1}{1.000.000 - 1} \left(\left(\sum_{j=1}^{1.000.000} (\Delta p'_j)^2 \right) - \frac{1}{1.000.000} \left(\sum_{j=1}^{1.000.000} \Delta p'_j \right)^2 \right) \tag{4.5}$$

ermittelt.

Zur Interpretation der Ergebnisse werden wiederum dieselben Grenzen wie in der Voruntersuchung verwendet. In Anlehnung an Cohen (1997, S. 157) beziehungsweise Fröhlich und Pieter (2009, S. 141) werden absolute Effektstärken ab 0,1 als bedeutsam eingestuft. Durch die Berechnung von d' lässt sich auch das Vorzeichen so interpretieren wie in der Vorstudie. $d' < 0$ deutet auf die Überlegenheit eines Donor-Limits von Eins hin. $d' > 0$ bedeutet, dass kein Donor-Limit von Vorteil ist.

4.2.2 Ergebnisse

Um die Einflussfaktoren hinsichtlich der Vorteilhaftigkeit eines Donor-Limits tiefgreifender zu untersuchen, werden im Folgenden die Ergebnisse der Simulationsstudie betrachtet. Die Präsentation der Ergebnisse erfolgt grafisch, nach einem einheitlichen Schema. Die Abbildungen 4.2 bis 4.19 bestehen jeweils aus vier Unterabbildungen. Diese Unterabbildungen sind so angeordnet, dass die Grafiken nebeneinander immer die Ergebnisse für eine unterschiedliche Merkmalsanzahl und die Grafiken untereinander immer die Ergebnisse für eine

unterschiedliche Objektanzahl wiedergeben. Innerhalb einer einzelnen Unterabbildung werden stets die ermittelten Effektstärken nach Cohen gegen den Anteil fehlender Werte abgetragen. Eine dieser Grafiken enthält drei Kurven, welche die zehn Stützstellen verbinden. Jede dieser Kurven stellt die Ergebnisse für eine bestimmte Korrelation der simulierten Daten dar. Dargestellt und analysiert werden demnach die Ergebnisse der Simulation für jede Faktorstufe des vollfaktoriellen Experimentaldesigns, um die in der Voruntersuchung identifizierten Wechselwirkungen näher zu untersuchen. Die kritischen Schranken für die Bedeutsamkeit von 0,1 beziehungsweise $-0{,}1$ sind als gestrichelte Linie eingezeichnet.

Im Vordergrund der Betrachtungen steht die Skalierung des Merkmals, das imputiert wurde. Hierfür werden die Ergebnisse zunächst separat für die Fälle des metrischen, ordinalen und nominalen Merkmals betrachtet. Im Anschluss erfolgt im letzten Unterabschnitt eine Darstellung der Wirkung der Skalenvariierung.

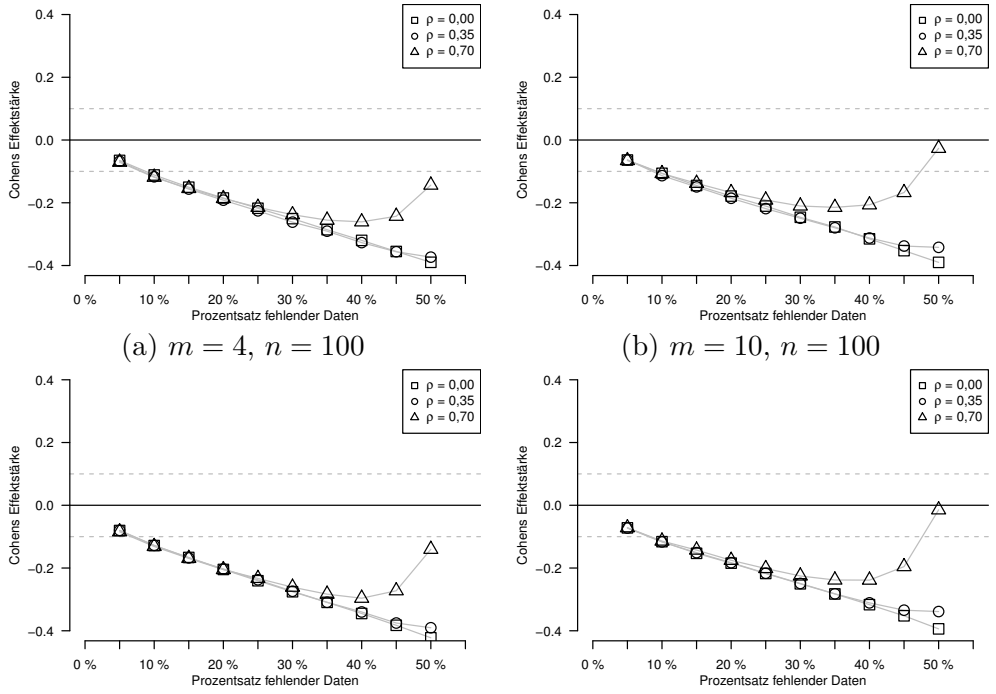
4.2.2.1 Auswirkungen bei metrischer Skalierung

Der folgende Unterabschnitt widmet sich der Betrachtung der Ergebnisse der Simulationsstudie für den Fall, dass ein metrisches Merkmal Datenausfall erfahren hat und imputiert wurde. Eine Beschreibung der Auswirkungen auf die Schätzung der Varianz des metrischen Merkmals und der Pearson-Korrelation zwischen dem imputierten Merkmal und einem weiteren metrischen Merkmal erfolgt getrennt nach Ausfallmechanismus. Die drei beschriebenen Ausfallmechanismen werden in aufsteigender Intensität, von *MCAR* bis *MAR 1:4*, diskutiert. Bei der Feststellung, welche Auswirkungen diese Intensivierung hat, wird Bezug auf die vorher dargestellten, milderen Fälle genommen.

MCAR-Ausfall

Die Auswirkung der Verwendung eines Donor-Limits auf die Schätzung der Varianz eines metrischen Merkmals ist im Falle eines *MCAR*-Ausfallmechanismus eindeutig. Die berechneten Effektstärken sind niemals positiv, daher führt die Verwendung des Donor-Limits grundsätzlich zu besseren Ergebnissen. In allen Unterabbildungen der Abbildung 4.2 ist qualitativ derselbe Kurvenverlauf zu erkennen. Der Vorteil, den die Verwendung eines Donor-Limits von Eins

bietet, steigt tendenziell mit dem Anteil der fehlenden Daten im zu imputierenden Merkmal. Dies ist durch die kleiner werdenden d' zu erkennen. Für einen Anteil an fehlenden Werten von mehr als 0,05 liegt der Vorteil fast immer über der Schwelle zur Bedeutsamkeit. Für eine hohe Korrelation, $\rho = 0{,}70$, lässt sich feststellen, dass die berechneten d' ein Minimum bei ca. 40% fehlender Werte erreichen. Ab diesem Punkt wird der Vorteil wieder kleiner. Für andere Korrelationen steigt der Vorteil monoton bis zu dem größten betrachteten Anteil an fehlenden Werten, 0,50. Die d' sind für größere ρ auch tendenziell größer.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

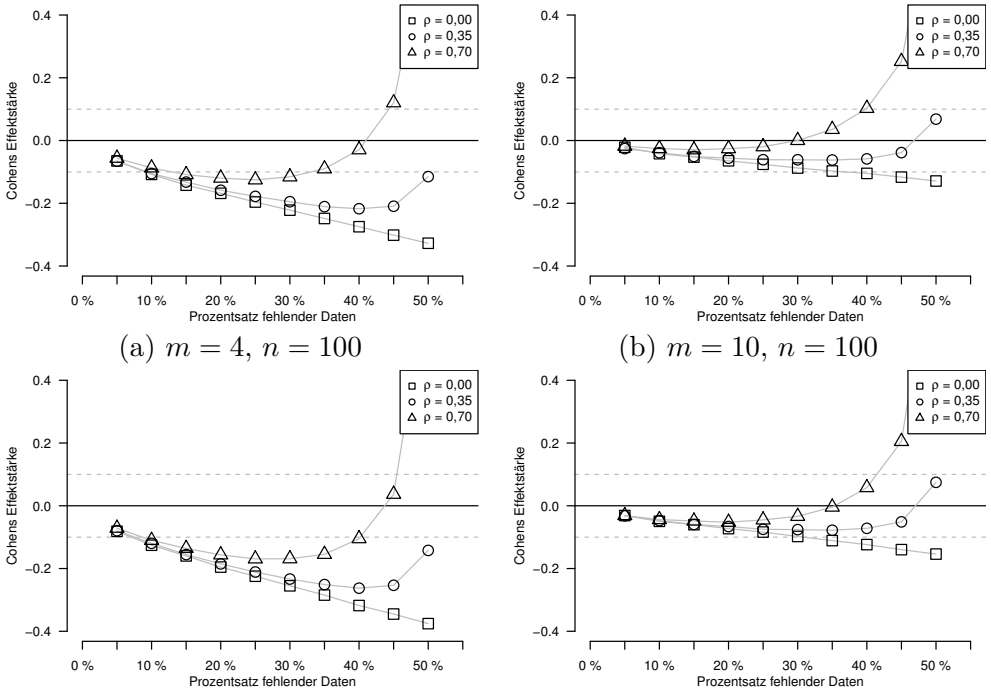
(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.2: Auswirkung von *MCAR*-Ausfall auf die Varianz eines metrisch skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Eine Erhöhung der Merkmalsanzahl führt zu einer betragsmäßigen Verringerung der Effektstärken, erkennbar an einer leichten Vertikalverschiebung der Kurven nach oben. Die Auswirkungen einer Erhöhung der Objektanzahl sind kaum ersichtlich. Leichte Vertikalverschiebungen der Kurven nach unten könnten auch Spreizungen zwischen den Kurven darstellen.

Die Auswirkung der Verwendung eines Donor-Limits auf die Schätzung der Korrelation zwischen dem imputierten und einem weiteren metrischen Merkmal ist im Falle eines *MCAR*-Ausfallmechanismus nicht eindeutig. Die be-

rechneten Effektstärken sind teilweise positiv, teilweise negativ und für einige Fälle deutlich unter der Schwelle zur Bedeutsamkeit. Alle Kurven in den vier Unterabbildungen 4.3a bis 4.3d weisen gewisse Ähnlichkeiten auf. Zunächst sinken die d' mit einem zunehmenden Anteil von fehlenden Werten. Danach steigen die d' für alle $\rho \neq 0{,}00$ wieder. Da auch die Minima für $\rho = 0{,}70$ und $\rho = 0{,}35$ unterschiedlich sind, kommt es bei größeren Anteilen an fehlenden Daten zu einer Spreizung der Kurvenverläufe. Während bei den zwei kleineren Korrelationen ein Donor-Limit nie wirklich nennenswert zum Nachteil für die Imputationsqualität wird, ist dies für $\rho = 0{,}70$ anders. Bei extremen Bedingungen von 45% beziehungsweise 50% fehlender Merkmalsausprägungen führt ein Donor-Limit zu einer erheblich schlechteren Imputationsqualität. Die Effektstärken bei 50% Ausfall in den Abbildungen 4.3a, 4.3b, 4.3c und 4.3d betragen 0,67; 0,82; 0,72 sowie 0,89 und sind somit deutlich als groß zu bezeichnen.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

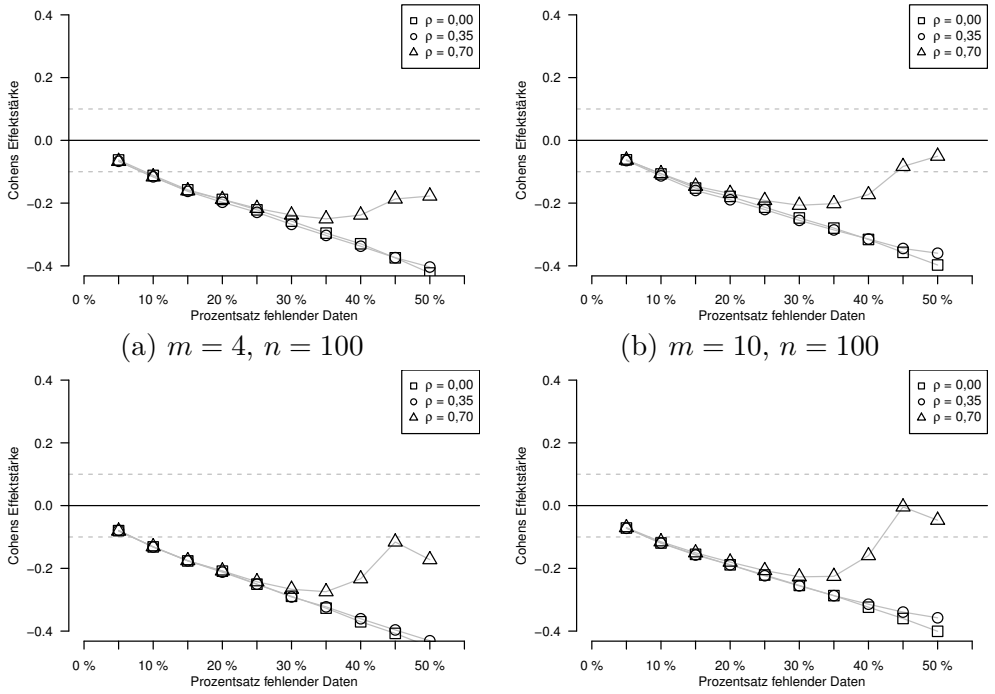
Abbildung 4.3: Auswirkung von *MCAR*-Ausfall auf die Korrelation bei metrisch skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Die Einbringung von mehr Merkmalen führt tendenziell zu einer Verschiebung der Kurven Richtung null. Eine Ausnahme stellen hier wieder die Extremfälle von 50% Datenausfall bei $\rho = 0{,}70$ dar. Hier sind die d' nicht näher

Null, sondern größer. Eine Erhöhung der Objektanzahl wirkt sich hingegen positiv auf die Vorteilhaftigkeit des Donor-Limits aus. Dies ist durch eine leichte Vertikalverschiebung der Kurven im Vergleich der Unterabbildungen 4.3a und 4.3c beziehungsweise 4.3b und 4.3d zu erkennen.

MAR 1:2-Ausfall

Der Effekt des Donor-Limits auf die Varianzschätzung bei einem metrischen Merkmal, das dem verschärften Ausfallmechanismus *MAR 1:2* ausgesetzt wurde, ist durchweg vorteilhaft. Dies ist an den durchgängig negativen d' Werten in der Abbildung 4.4 zu erkennen. Der Verlauf der Kurven ist in allen vier Unterabbildungen einheitlich. Die Effektstärken beginnen innerhalb der Trivialitätsgrenze für 5% Datenausfall und fallen mit zunehmendem Datenausfall, für $\rho = 0{,}00$ und $\rho = 0{,}35$ stets, bis zu $d' \approx 0{,}40$ weiter ab. Für $\rho = 0{,}70$ ergibt sich ca. bei 30–35% Datenausfall ein Minimum und bei 45% ein Maximum im Kurvenverlauf. Bei $m = 10$ und 45% oder mehr Datenausfall ist der Vorteil der Hot-Deck-Imputation mit einem Donor-Limit nur noch trivial.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

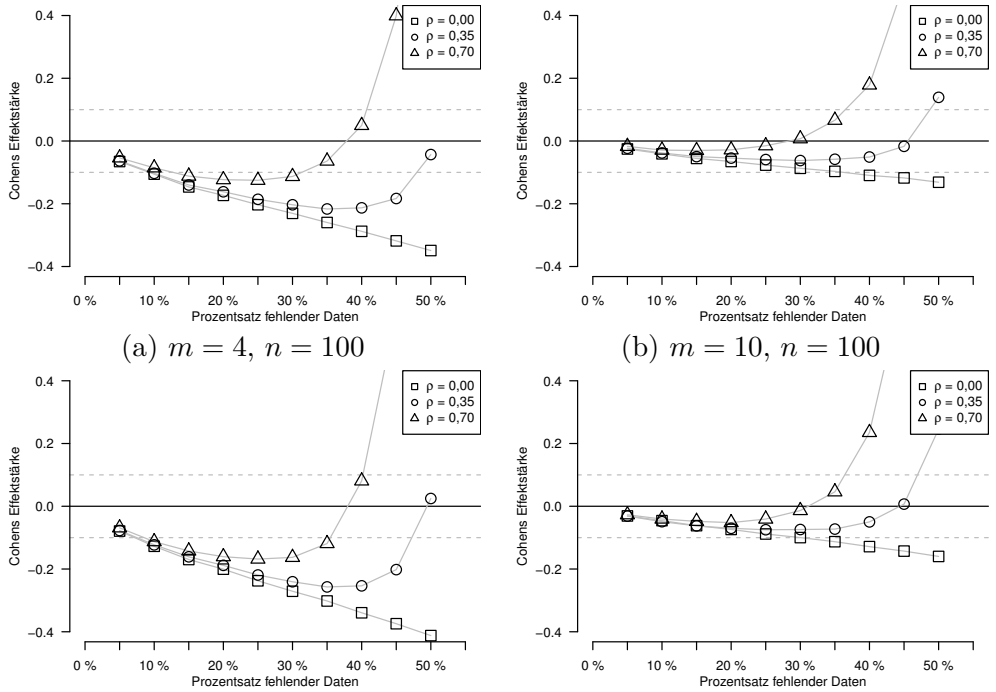
Abbildung 4.4: Auswirkung von *MAR 1:2*-Ausfall auf die Varianz eines metrisch skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Eine Reduzierung der Merkmalsanzahl führt zu einer geringen Reduzierung der d', welches auf einen größeren Vorteil der Verwendung eines Donor-Limits hinweist. Eine Verringerung der Objektzahl in den simulierten Datenmatrizen führt tendenziell zu einer Stauchung der Kurvenverläufe zueinander.

Grundsätzlich ergeben sich im direkten Vergleich zu den Varianzschätzungen bei dem *MCAR*-Datenausfall nur geringe Unterschiede. Für die meisten Fälle ist bei dem *MAR 1:2* ein Donor-Limit deutlich mehr vorteilhaft als beim *MCAR*-Datenausfall. Lediglich bei einer extremen Anzahl an fehlenden Werten und dem $\rho = 0{,}70$ sind zudem im Verlauf der d' qualitative Unterschiede vorhanden. So sind hier, beim intensiveren Ausfallmechanismus, bei zehn Merkmalen bereits ab 45% die Unterschiede in der Imputationsqualität der Verfahren trivial.

Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits auf die Schätzung der Korrelation zwischen dem imputierten und einem weiteren metrischen Merkmal ist im Falle des intensiveren *MAR 1:2*-Ausfallmechanismus nicht eindeutig. Effektstärken sind in Abhängigkeit der Anzahl an fehlenden Werten, der vorherrschenden Korrelation innerhalb der simulierten Datenmatrizen und der Merkmalsanzahl entweder negativ oder positiv. Vorwiegend sind die berechneten d' jedoch negativ (101 der 120 in Abbildung 4.5 dargestellten Fälle), so dass festgestellt werden kann, dass auch hier eine Beschränkung der Spenderverwendung von Vorteil ist. Lediglich in 11 Fällen, wo der Datenausfall sehr hoch und auch ρ mittel oder groß ist, ist es für die Schätzung der Korrelation vorteilhaft, dass die Spendernutzung nicht eingeschränkt wird. Hier ist zugleich der Vorteil extrem groß. Beispielsweise betragen in den Unterabbildungen 4.5a, 4.5b, 4.5c und 4.5d die d' bei $\rho = 0{,}70$ und einem Datenausfall von 50% 1,20; 1,36; 1,82 und 2,01.

Bei diesem Ausfallmechanismus ist der markante, durchgebogene Kurvenverlauf nicht nur bei $\rho = 0{,}70$, sondern auch bei $\rho = 0{,}35$ in allen Unterabbildungen zu erkennen. Die Erreichung des Minimums wird hier zudem durch die Anzahl der Merkmale beeinflusst. Für $m = 10$ ist dies jedoch nicht weiter interessant, da hier die meisten d' trivial sind. Mit $m = 4$ sind die Effekte deutlich interessanter. Die Kurven deuten auf einen merklichen Vorteil des Donor-Limits hin. Lediglich bei $\rho = 0{,}70$ und einem Datenausfall von mehr als 40% bietet eine Imputation ohne Donor-Limit einen Vorteil.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.5: Auswirkung von *MAR 1:2*-Ausfall auf die Korrelation bei metrisch skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

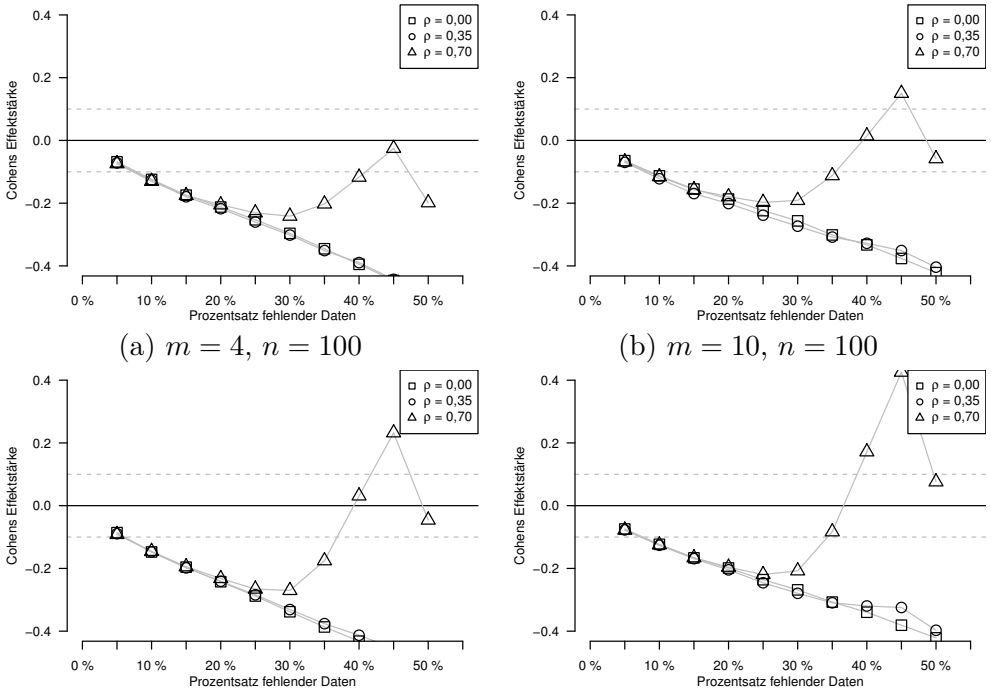
Im direkten Vergleich zum *MCAR*-Ausfallmechanismus kann festgehalten werden, dass für $\rho = 0{,}00$ keine Unterschiede in den d' identifizierbar sind. Für $\rho = 0{,}35$ und $\rho = 0{,}70$ sind Unterschiede vorhanden, jedoch lediglich für die größeren Anteile an fehlenden Werten. Diese Unterschiede sind im direkten Vergleich der Abbildungen 4.3 und 4.5 für einen Anteil fehlender Werte größer gleich 0,40 anhand eines steileren Anstiegs der Kurven zu erkennen.

MAR 1:4-Ausfall

Im Falle des intensivsten der betrachteten Ausfallmechanismen ist die Vorteilhaftigkeit eines Donor-Limits für eine Varianzschätzung bei einem metrischen Merkmal nicht gänzlich eindeutig. In nahezu allen simulierten Fällen ist ein Donor-Limit nicht von Nachteil. 116 der in Abbildung 4.6 dargestellten 200 Fälle weisen ein d' kleiner als 0,1 auf. In lediglich 4 Fällen führt eine Hot-Deck-Imputation ohne Donor-Limit zu einer deutlich besseren Imputationsqualität.

Der Kurvenverlauf beginnt in den Unterabbildungen 4.6a bis 4.6d immer gleich. Zunächst sind alle d' bei 5% Datenausfall leicht negativ und nehmen mit einem zunehmenden Anteil fehlender Werte weiter ab. Lediglich für $\rho = 0{,}70$

ändert sich dieser Verlauf ab 30% fehlender Daten. Hier ist es vergleichsweise besser, kein Donor-Limit zu verwenden, bis ein Maximum bei einem Anteil fehlender Werte von 0,45 erreicht wird. Nach diesem Maximum erfolgt ein steiler Abfall in den d' Werten.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.6: Auswirkung von *MAR 1:4*-Ausfall auf die Varianz eines metrisch skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Grundsätzlich ergeben sich im direkten Vergleich zu den Varianzschätzungen bei dem *MAR 1:2*-Ausfallmechanismus auch nur geringe Unterschiede. Wird die Progression in Summe betrachtet, so sind die Unterschiede zum *MCAR*-Ausfallmechanismus jedoch bereits merklich und zeigen die Auswirkungen der Intensivierung des Ausfallmechanismus. Bei dem geringen und mittleren ρ sind die d' über alle Fälle hinweg konsistent negativ. Eine Intensivierung des Ausfallmechanismus führt in diesen Fällen jeweils nur zu einer Ausweitung der Vorteilhaftigkeit eines Donor-Limits. Für $\rho = 0{,}70$ sind die d' nicht über alle Fälle hinweg negativ. Hier hat die Intensivierung des Ausfallmechanismus einen etwas heterogenen Einfluss. Grundsätzlich werden die Effektgrößen größer, so dass diese in den extremsten der betrachteten Fällen für eine Imputation ohne Donor-Limit sprechen. Jedoch entsteht auch ein Knick

im Kurvenverlauf. Hier (bei 50% fehlender Daten) ist die simulierte Situation derart extrem, dass sich die Imputationsqualität der Hot-Deck-Verfahren mit und ohne Donor-Limit wieder angleicht.

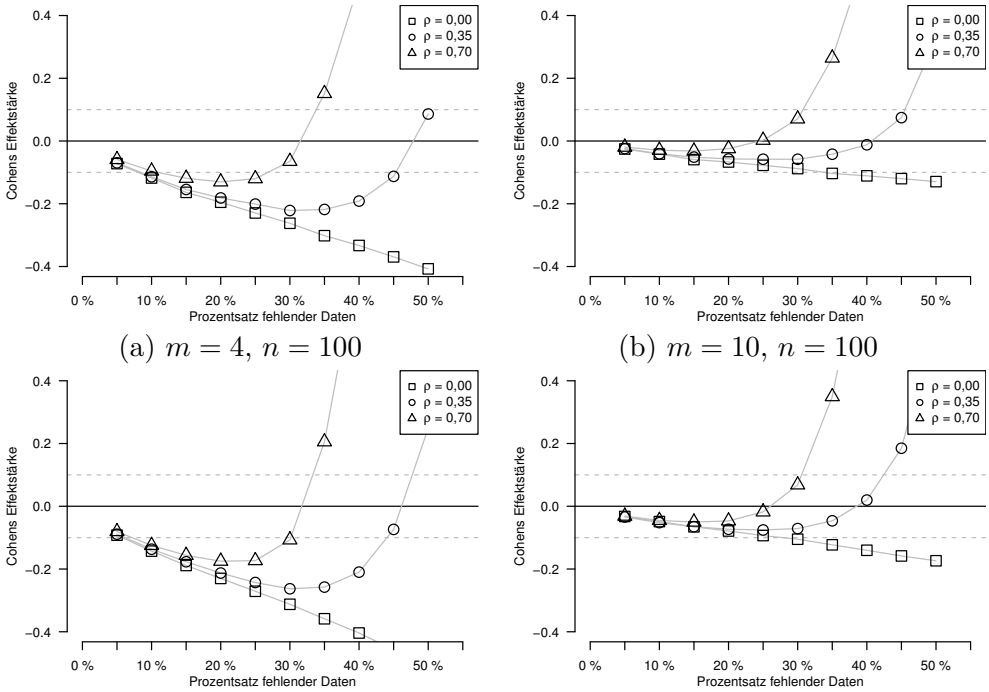
Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits bei der Schätzung der Korrelation zwischen dem imputierten und einem weiteren metrischen Merkmal ist auch im Falle des stärksten Ausfallmechanismus, *MAR 1:4*, nicht eindeutig. Die berechneten d' sind insbesondere in Abhängigkeit der vorherrschenden Korrelation innerhalb der simulierten Datenmatrizen entweder positiv oder negativ. In 50 von 120 Fällen ist die Imputation mit Donor-Limit überlegen, in 20 von 120 Fällen ist eine Imputation ohne Donor-Limit besser und in den verbleibenden 50 Fällen sind die Effektstärken betragsmäßig trivial. Somit ist auch in den hiesig betrachteten Fällen meist ein Donor-Limit für die Imputationsqualität nicht von Nachteil und am häufigsten vorteilhaft.

Eine Vorteilhaftigkeit der Hot-Deck-Imputation ohne Donor-Limit tritt meist wieder unter den extremeren der betrachteten Fälle auf. So entstehen $d' \geq 0{,}1$ insbesondere bei den größeren ρ und einem größeren Anteil fehlender Werte. Diese Effekte, und der damit assoziierte Vorteil bei der Imputationsqualität, werden, in Abhängigkeit des Datenausfalls, zu groß, um in Abbildung 4.7 dargestellt zu werden. Beispielsweise entspricht das d' für ein $\rho = 0{,}70$ und einer 200×10 simulierten Matrix bei einem Datenausfall von 50% einem Wert[4] von ca. 3,82.

Im Allgemeinen führt eine Erhöhung der Merkmalsanzahl zu einer Erhöhung der Effektstärken, erkennbar an einer Vertikalverschiebung der Kurven in Abbildung 4.7 nach oben. Die Auswirkungen einer Erhöhung der Objektanzahl sind bei einer geringeren Merkmalsanzahl besser zu erkennen als bei einer größeren. Beim Vorhandensein von mehr Objekten ist auch eher ein Donor-Limit von Vorteil. Dies ist besonders ersichtlich im Vergleich der Kurven für $\rho = 0{,}00$ beziehungsweise $\rho = 0{,}70$ in den Unterabbildungen 4.7a und 4.7c. Insgesamt ist die Auswirkung eher klein und wird deutlich von der Auswirkung einer Merkmalserhöhung überlagert.

---

[4] Dies ist das größte d', welches für die Schätzung der Korrelation zwischen dem imputierten und einem weiteren metrischen Merkmal berechnet wurde.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.7: Auswirkung von *MAR 1:4*-Ausfall auf die Korrelation bei metrisch skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Auch bei der Korrelationsschätzung ergeben sich im Allgemeinen und im direkten Vergleich zu dem *MAR 1:2*-Ausfallmechanismus geringere Unterschiede als zum *MCAR*-Ausfallmechanismus. In den simulierten Fällen bewirkt der Übergang von *MCAR* zu *MAR 1:2* zu *MAR 1:4* tendenziell eine Verstärkung der vorhandenen Effekte auf die berechneten d'. Nur in seltenen Fällen wird ein Vorzeichenwechsel durch die Variation des Ausfallmechanismus herbeigeführt. Diese seltenen Fälle sind zudem durch die extremen Ausprägungen anderer Faktoren charakterisiert[5].

4.2.2.2 Auswirkungen bei ordinaler Skalierung

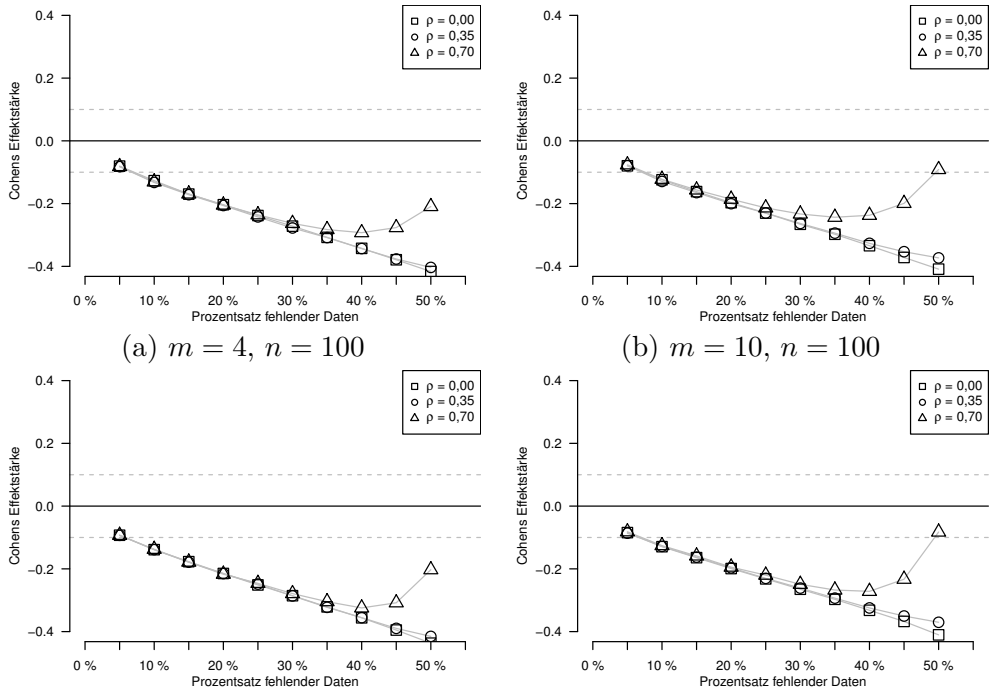
Im folgendem Unterabschnitt werden die Ergebnisse der Simulationsstudie für den Fall, dass bei einem ordinalen Merkmal Datenausfall zu verzeichnen ist, betrachtet. Beschrieben werden die Auswirkungen auf die Schätzung der Quartilsdifferenz des ordinalen Merkmals und der Spearman-Rangkorrelation zwischen dem imputierten Merkmal und einem weiteren ordinalen Merkmal, getrennt

[5] Beispielsweise, wenn 50% der Werte fehlen.

nach Ausfallmechanismus. Die Behandlung der drei beschriebenen Ausfallmechanismen erfolgt in aufsteigender Intensität, von *MCAR* bis *MAR 1:4*. Bei der Feststellung, welche Auswirkungen diese Intensivierung hat, wird Bezug auf die vorher dargestellten, milderen Fälle genommen.

MCAR-Ausfall

Die Auswirkung der Verwendung eines Donor-Limits auf die Schätzung der Quartilsdifferenz eines ordinalen Merkmals ist im Falle eines *MCAR*-Ausfallmechanismus eindeutig. Die berechneten Effektstärken sind niemals positiv, daher führt die Verwendung des Donor-Limits grundsätzlich zu besseren Ergebnissen. Lediglich in 14 Fällen sind diese Effektstärken als nicht bedeutsam einzustufen. Die Unterabbildungen der Abbildung 4.8 zeigen alle ähnliche Kurvenverläufe.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

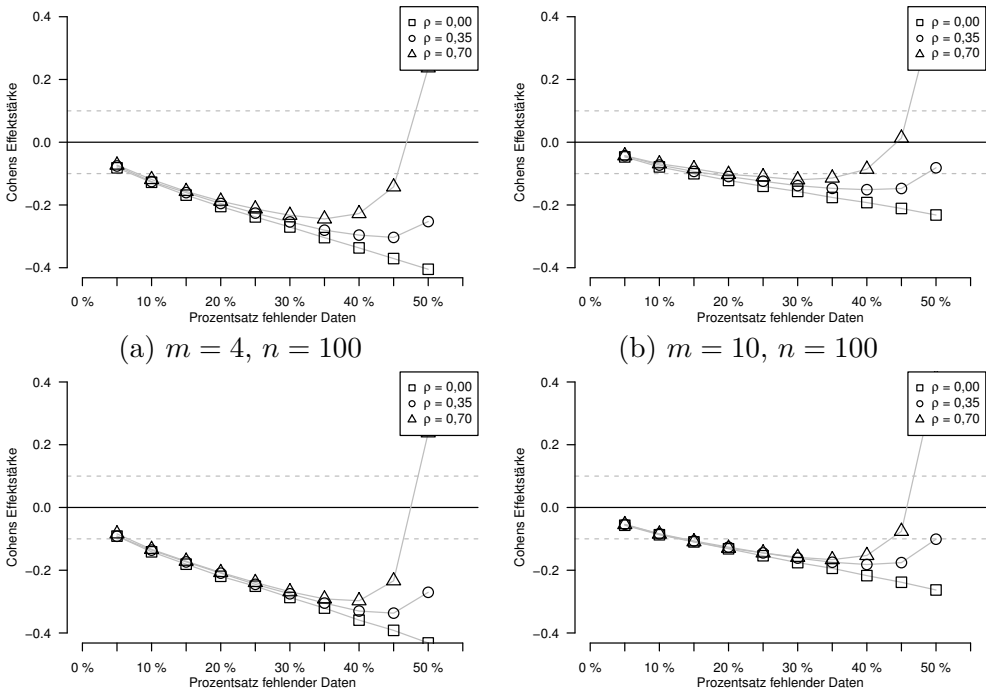
(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.8: Auswirkung von *MCAR*-Ausfall auf die Quartilsdifferenz eines ordinal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Die Vorteilhaftigkeit des Verfahrens mit Donor-Limit steigt tendenziell mit dem Anteil der fehlenden Daten im zu imputierenden Merkmal. Dies ist anhand der kleiner werdenden d' zu erkennen. Für einen Anteil an fehlenden Werten

von mehr als 0,05 liegt der Vorteil bis auf zwei Fälle[6] über der Schwelle zur Bedeutsamkeit. Für eine hohe Korrelation, $\rho = 0{,}70$, lässt sich feststellen, dass die berechneten d' ein Minimum bei ca. 40% fehlender Werte erreichen. Ab diesem Punkt wird der Vorteil wieder kleiner. Bei den zwei anderen ρ steigt der Vorteil monoton bis zu 50% fehlender Werte. Die d' sind für größere ρ auch tendenziell größer.

Werden die Ergebnisse nach Merkmalsanzahl verglichen, ist erkennbar, dass die Effektstärken bei den Fällen mit einer größeren Merkmalsanzahl immer kleiner sind als bei den Fällen mit einer kleineren Merkmalsanzahl. Eine Variation der Objektanzahl erzeugt kaum eine Veränderung der d'.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.9: Auswirkung von *MCAR*-Ausfall auf die Rangkorrelation bei ordinal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Soll die Rangkorrelation zwischen dem imputierten und einem weiteren ordinalen Merkmal geschätzt werden, sind die Auswirkungen eines Donor-Limits bei der Hot-Deck-Imputation von *MCAR*-fehlenden Daten nicht vollständig eindeutig. Die berechneten Effektstärken sind nahezu immer negativ und be-

[6] Diese zwei Fälle sind $\rho = 0{,}70$, $m = 9$ und 50% fehlende Werte. Die zugehörigen d' betragen $-0{,}02$ für $n = 100$ und -0.01 für $n = 200$.

deutsam. In lediglich vier Fällen[7] sind die $d' \geq 0{,}1$, welches gegen ein Donor-Limit sprechen würde.

Alle Kurven in den vier Unterabbildungen 4.9a bis 4.9d weisen gewisse Ähnlichkeiten auf. Zunächst sinken die d' mit einem zunehmenden Anteil von fehlenden Werten. Danach steigen die d' für alle $\rho \neq 0{,}00$ wieder. Da auch die Minima für $\rho = 0{,}70$ und $\rho = 0{,}35$ unterschiedlich sind, kommt es bei größeren Anteilen an fehlenden Daten zu einer Spreizung der Kurvenverläufe.

Die Einbringung von mehr Merkmalen führt teilweise zu einer erheblichen Vertikalverschiebung der Kurven nach oben. Die Vorteile, die ein Donor-Limit für die Imputationsqualität bietet, reduzieren sich hierdurch. Eine Erhöhung der Objektanzahl wirkt sich hingegen positiv auf die Vorteilhaftigkeit des Donor-Limits aus. Dies ist an einer Verschiebung der Kurven im Vergleich der Unterabbildungen 4.9a und 4.9c beziehungsweise 4.9b und 4.9d zu erkennen.

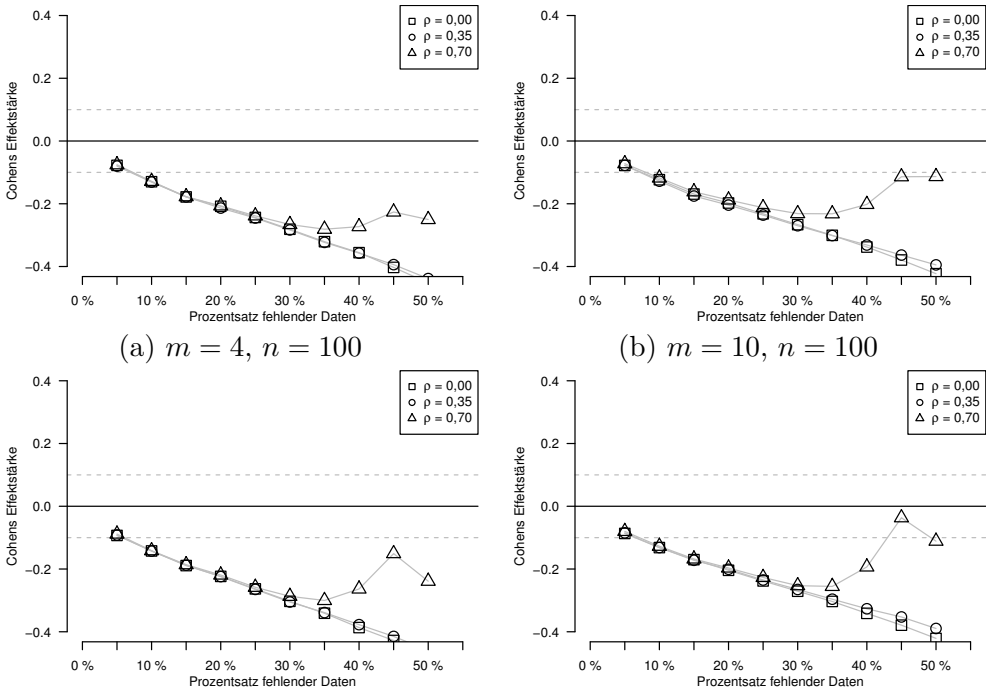
***MAR 1:2*-Ausfall**

Der Effekt des Donor-Limits auf die Quartilsdifferenzsschätzung bei einem ordinalen Merkmal, das dem verschärften Ausfallmechanismus *MAR 1:2* ausgesetzt wurde, ist durchweg vorteilhaft. Dies ist an den durchgängig negativen d'-Werten in der Abbildung 4.10 zu erkennen. Der Verlauf der Kurven ist in allen vier Unterabbildungen einheitlich. Die Effektstärken beginnen innerhalb der Trivialitätsgrenze für 5% Datenausfall und fallen bis zu einem Datenausfall von 35% ab. Während die d' für größere Anteile fehlender Daten bei kleinem und mittlerem ρ weiter abfallen, ergibt sich für das größte betrachtete ρ hier ein Minimum. Im weiteren Verlauf von $\rho = 0{,}70$ entsteht dann bei einem Anteil fehlender Werte von 0,45 ein Maximum.

Eine Erhöhung der Objektanzahl führt kaum zu einer Veränderung in den d'. Diese Veränderungen erhöhen meist den Vorteil für das Verfahren mit Donor-Limit, mit Ausnahme der rechten Extreme des Verlaufs bei $\rho = 0{,}70$. In diesem Fall verursacht eine Erhöhung der Objektanzahl eine Verringerung des Donor-Limit-Vorteils. Die Erhöhung der Merkmalsanzahl führt stets zu einer Erhöhung der d'. Dies weist auf eine Reduzierung des Vorteils, mit Donor-Limit zu imputieren, hin.

---

[7] $\rho = 0{,}70$ und 50% fehlende Daten.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

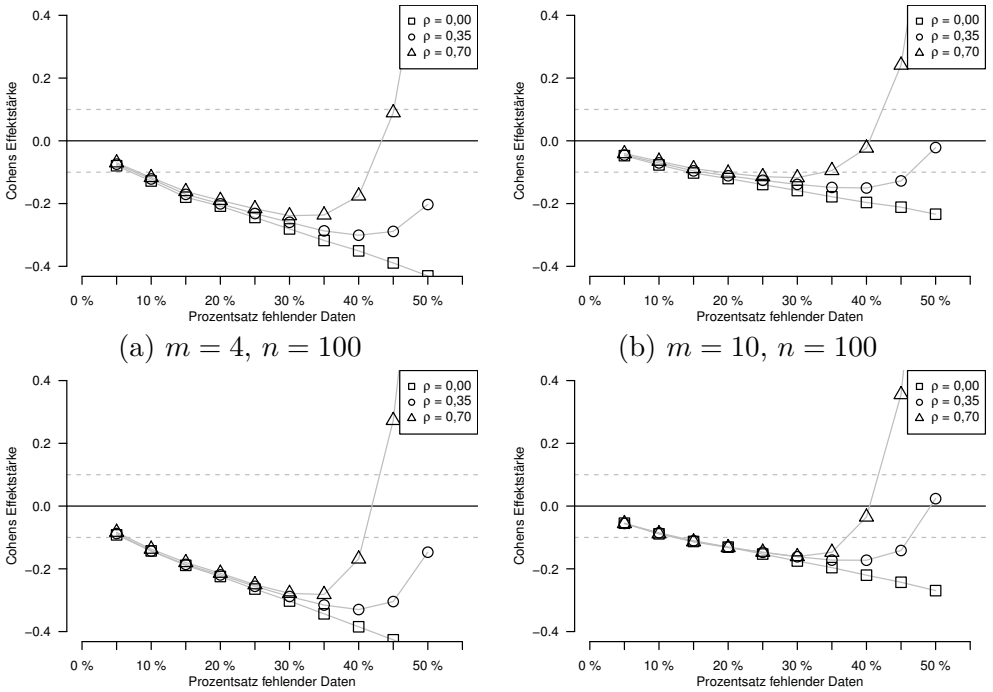
(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.10: Auswirkung von *MAR 1:2*-Ausfall auf die Quartilsdifferenz eines ordinal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Werden für die Schätzung dieses Parameters die Auswirkungen des *MCAR*- und *MAR 1:2*-Ausfallmechanismus verglichen, sind nur geringe Veränderungen in den Effektstärken erkennbar. Am auffälligsten sind Veränderungen im Kurvenverlauf von $\rho = 0{,}70$. Hier wird der parabelähnliche Verlauf markanter. Auch der Knick bei 45% fehlender Daten ist beim *MCAR*-Ausfallmechanismus nicht vorhanden. Abseits dieser Randerscheinungen erhöht sich die Vorteilhaftigkeit der Hot-Deck-Imputation mit Donor-Limit durch die Intensivierung des Ausfallmechanismus nur leicht.

Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits bezüglich der Schätzung der Rangkorrelation zwischen dem imputierten und einem weiteren ordinalen Merkmal sind im Falle des intensiveren *MAR 1:2*-Ausfallmechanismus nicht eindeutig. Effektstärken sind insbesondere in Abhängigkeit der Anzahl an fehlenden Werten und der vorherrschenden Korrelation innerhalb der simulierten Datenmatrizen entweder negativ oder positiv. Jedoch sind lediglich in sieben Fällen die berechneten $d' \geq 0{,}1$; daher muss festgestellt werden, dass hier die Anwendung des Donor-Limits fast

immer nicht von Nachteil für die Rangkorrelationsschätzung ist. Wie den Unterabbildungen 4.11a bis 4.11d entnommen werden kann, ist kein Donor-Limit zu verwenden, nur unter den extremen Umständen von 45 bis 50 Prozent fehlender Werte bei $\rho = 0{,}70$ empfehlenswert[8]. In den verbleibenden 116 Fällen ist ein Donor-Limit in 26 Fällen nicht wesentlich schlechter oder besser und in 87 Fällen sogar deutlich besser.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.11: Auswirkung von *MAR 1:2*-Ausfall auf die Rangkorrelation bei ordinal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Wie auch bereits beim *MCAR*-Datenausfall ist hier für jedes $\rho > 0{,}00$ der markante, durchgebogene Kurvenverlauf in der Abbildung 4.11 zu erkennen. Die Durchbiegung ist bei der kleineren Merkmalsanzahl deutlich größer als bei der größeren. Bei $m = 10$ sprechen die Effektgrößen im Wesentlichen für ein Donor-Limit, wenn der Anteil fehlender Werte zwischen 15 und 35% beträgt. Für kleinere und meist auch größere Anteile an fehlenden Werten sind die d' zwar auch negativ, aber unbedeutend. Eine Erhöhung der Objektanzahl scheint vorhandene Effekte zu intensivieren, die d' werden tendenziell mit steigender

[8] Ausnahme hierbei ist die Konstellation $m = 3$, $n = 100$, $\rho = 0{,}70$ und 45% fehlender Werte. Hier ist $d' = 0{,}089$ innerhalb der Trivialitätsgrenzen.

Objektanzahl betragsmäßig größer.

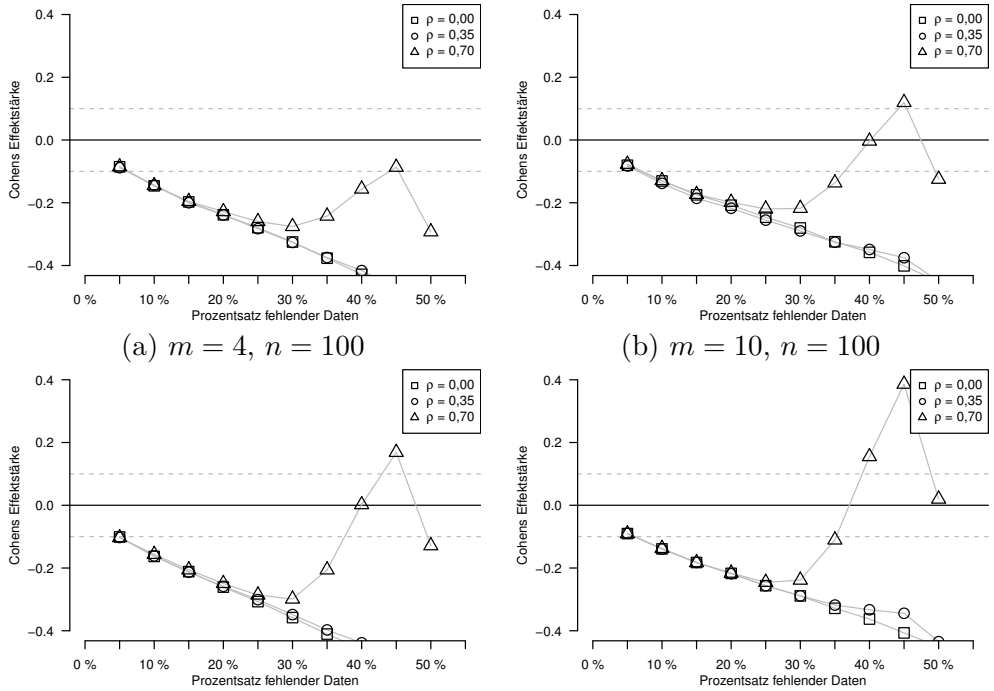
Im direkten Vergleich zu den Ergebnissen beim *MCAR*-Ausfallmechanismus ergeben sich kaum qualitative Unterschiede in den Kurvenverläufen. Zu erkennen ist lediglich, dass beim *MAR 1:2*-Ausfallmechanismus ein Donor-Limit deutlich stärker von Vorteil für die Rangkorrelationsschätzung ist. Die d' in der Abbildung 4.11 werden mit steigendem Anteil fehlender Werte deutlich stärker negativ und bleiben länger und weiter unterhalb von $-0{,}1$ im Vergleich zur Abbildung 4.9.

***MAR 1:4*-Ausfall**

Werden jene simulierten Fälle betrachtet, bei denen eine Schätzung der Quartilsdifferenz für ein ordinales Merkmal unter dem intensivsten Ausfallmechanismus stattfindet, so ist eine Hot-Deck-Imputation mit einem Donor-Limit nicht immer die beste Wahl. Zwar ist es für $\rho \leq 0{,}35$ immer von Vorteil ein Donor-Limit zu verwenden, aber bei $\rho = 0{,}70$ kann es in Fällen mit einem extremen Datenausfall dazu kommen, dass eine Hot-Deck-Imputation ohne Begrenzung besser ist. Ein nennenswerter Vorteil wird jedoch nur selten erreicht. Er wird insbesondere dann erzielt, wenn der Anteil fehlender Werte 45% beträgt.

Der Kurvenverlauf ist bei allen Unterabbildungen 4.12a bis 4.12d sehr ähnlich. Alle Kurven beginnen immer bei $d' \approx -0{,}1$ und fallen zunächst mit einem zunehmenden Anteil fehlender Werte weiter ab. Während dies sich für $\rho = 0{,}00$ und $\rho = 0{,}35$ bis zu einem Anteil fehlender Werte von 50% fortsetzt, wird für $\rho = 0{,}70$ bei einem Anteil fehlender Werte von ca. 0,3 ein Minimum erreicht. Der Betrag der zugehörigen d' an den Minima hängt, ähnlich wie die Anstiege danach, von n und m ab. Eine Erhöhung der Objekt- und Merkmalsanzahl führt hier zu einer betragsmäßigen Erhöhung der d' und der Anstiege. Zusätzlich weist jede der vier Kurven ein weiteres Maximum bei 45% fehlenden Werten auf.

Werden die Quartilsdifferenzsschätzungen über die betrachteten Ausfallmechanismen hinweg verglichen, so ist grundsätzlich eine Intensivierung der Effektstärken festzustellen. Ist für eine bestimmte Faktorstufenkombination bei dem *MCAR*-Ausfallmechanismus ein Donor-Limit von Vorteil, so ist dies grundsätzlich unter dem *MAR 1:4*-Ausfallmechanismus auch deutlich besser.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

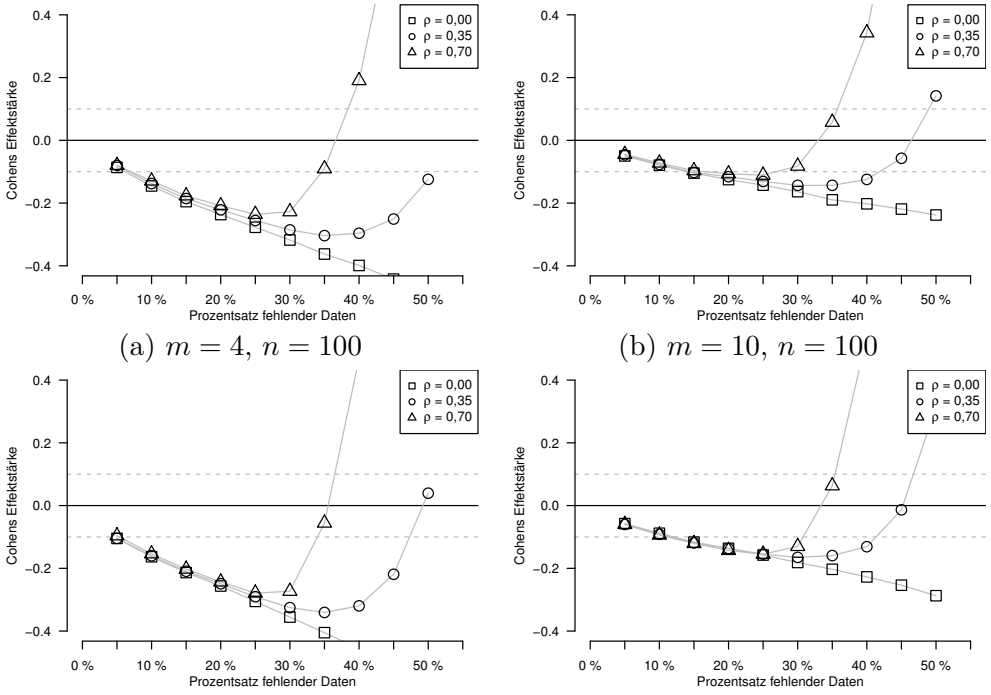
Abbildung 4.12: Auswirkung von *MAR 1:4*-Ausfall auf die Quartilsdifferenz eines ordinal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Ausnahme hierfür stellen einige der betrachteten Fälle dar, wenn die Korrelation innerhalb der simulierten Datenmatrix $\rho = 0{,}70$ beträgt. Hier ist – abweichend von der vorherigen Aussage – festzustellen, dass insbesondere der Anstieg nach dem Minimum durch eine Intensivierung des Ausfallmechanismus erhöht wird. So treten bei dem *MAR 1:4*-Ausfallmechanismus in Randfällen d' auf, die andeuten, dass eine Hot-Deck-Imputation ohne Donor-Limit für die Parameterschätzung von Vorteil ist.

Betrachtet man die Auswirkungen des intensivsten Ausfallmechanismus auf die Vorteilhaftigkeit des Donor-Limits für die Schätzung der Rangkorrelation zwischen dem imputierten und einem weiteren ordinalen Merkmal, so ist keine undifferenzierte Aussage möglich. In einem großen Anteil (103 von 120) der betrachteten Fälle ist die Imputation mit einem Donor-Limit grundsätzlich von Vorteil ($d' < 0$). Bei 81 dieser Fälle ist der Vorteil sogar bedeutsam ($d' \leq -0{,}1$). Gegen die Verwendung eines Donor-Limits sprechen demnach lediglich 17 Fälle ($d' > 0$), von denen 14 Fälle bedeutsame Unterschiede aufweisen ($d' \geq 0{,}1$). Diese Unterschiede treten zumeist bei höheren Korrelationen innerhalb der

simulierten Datenmatrizen und höheren Anteilen an fehlenden Werten auf.

Grundsätzlich beginnen alle Kurvenverläufe in der Abbildung 4.13 ähnlich. Bei 5% fehlenden Daten sind alle leicht negativ und werden mit wachsendem Anteil fehlender Werte zunehmend negativ bis ca. 25% fehlender Daten. Nach diesem Punkt ändert sich der Anstieg der Kurven in Abhängigkeit von ρ. Je größer ρ, desto größer die Veränderung im Anstieg. Bei $\rho = 0{,}00$ werden die berechneten d' weiter negativ. Für $\rho = 0{,}35$ biegt sich die Kurve in einem parabelähnlichen Verlauf leicht zunehmend nach oben. Mit $\rho = 0{,}70$ ist die Veränderung im Anstieg am größten, hier findet eine deutlich drastischere Durchbiegung der Kurve statt. Am Ende dieser Durchbiegungen (ab 40% fehlender Werte bei $\rho = 0{,}70$) ist es deutlich von Vorteil, eine Hot-Deck-Imputation ohne Donor-Limit durchzuführen. Hier treten d' von bis zu 3,06 bei $n = 200$ und $m = 10$ auf.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.13: Auswirkung von *MAR 1:4*-Ausfall auf die Rangkorrelation bei ordinal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Werden die im Falle des *MAR 1:4*-Ausfallmechanismus berechneten Ergebnisse mit denen unter weniger intensiven Ausfallmechanismen verglichen, kann eine, größtenteils kleine, Verstärkung der Effekte festgestellt werden. Fälle, bei

denen ein Donor-Limit für die Rangkorrelationsschätzung von Vorteil ist, wird tendenziell der Vorteil mit einer Intensivierung des Ausfallmechanismus größer. Ebenso wird in jenen Extremfällen, bei denen ein Donor-Limit von Nachteil ist, dieser Nachteil größer. Auch sind es diese Randfälle, bei denen die Intensivierung des Ausfallmechanismus zu den größten Veränderungen führt.

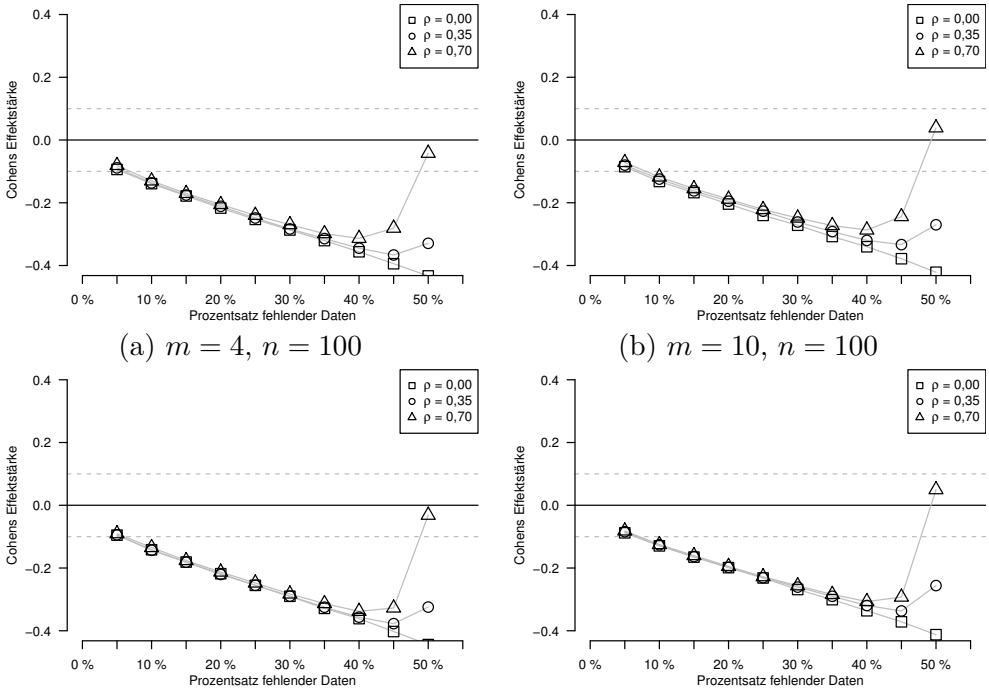
4.2.2.3 Auswirkungen bei nominaler Skalierung

Im nachfolgenden Unterabschnitt erfolgt eine Präsentation der Ergebnisse für den Fall, dass bei einem nominalen Merkmal fehlende Werte zu verzeichnen sind. Betrachtet werden die Auswirkungen auf eine Schätzung der Ausprägungshäufigkeit einer einzelnen Ausprägung des nominalen Merkmals und eine Schätzung des normierten Kontingenzkoeffizienten zwischen dem imputierten Merkmal und einem weiteren nominalen Merkmal. Die Ergebnisse werden getrennt nach den drei Ausfallmechanismen *MCAR*, *MAR 1:2* und *MAR 1:4* behandelt.

***MCAR*-Ausfall**

Soll die Ausprägungshäufigkeit eines nominalen Merkmals nach der Imputation von *MCAR*-fehlenden Daten geschätzt werden, ist die Empfehlung zur Nutzung eines Donor-Limits eindeutig. Die Verwendung eines Donor-Limits führt in keinem der betrachteten Fälle zu einem bedeutsamen Nachteil. Im Gegenteil, in allen – mit Ausnahme von zwei Fällen – ist ein Donor-Limit von Vorteil und in 104 Fällen sogar von erheblichem Vorteil ($d' \leq -0{,}1$).

Einflussreichster Faktor ist der Anteil fehlender Werte. Bis zu einem Anteil fehlender Werte von 0,4 nehmen die in Abbildung 4.14 dargestellten Kurven monoton ab. Erst bei 45% fehlender Werte oder mehr ist bei $\rho \geq 0{,}35$ ein Wechsel im Vorzeichen des Anstiegs zu verzeichnen. Die Anzahl der Objekte und Merkmale in der simulierten Datenmatrix haben beide einen vernachlässigbar kleinen Einfluss auf die Ergebnisse.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

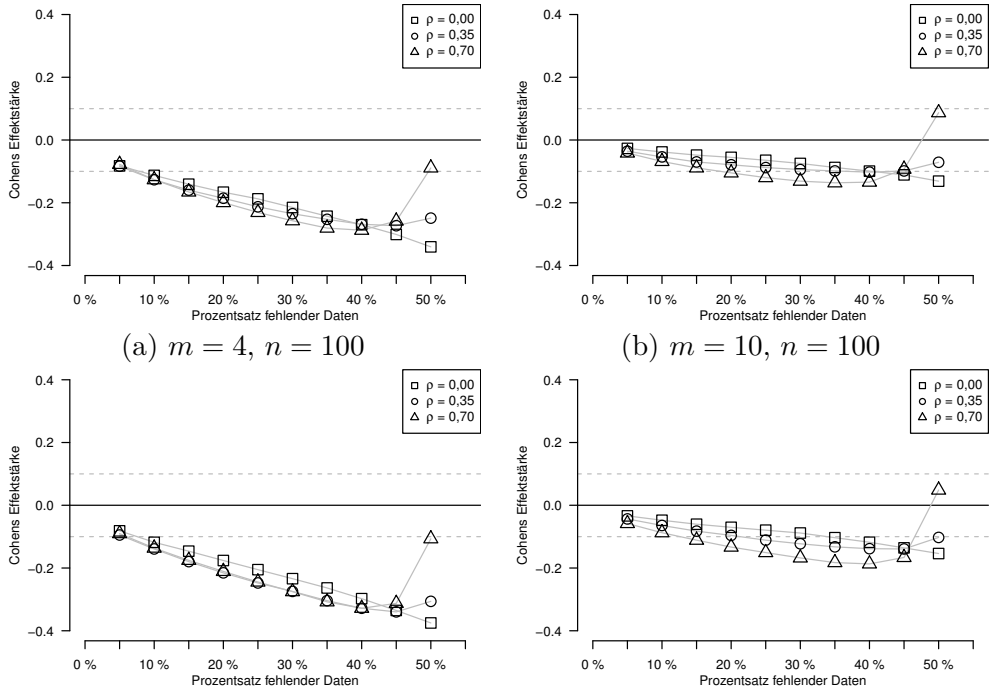
(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.14: Auswirkung von *MCAR*-Ausfall auf die Ausprägungshäufigkeit eines nominal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits bei Schätzung des normierten Kontingenzkoeffizenten zwischen dem imputierten und einem weiteren nominalen Merkmal ist im Falle des *MCAR*-Ausfallmechanismus eindeutig. In keinem der simulierten Fälle ist ein Donor-Limit erheblich von Nachteil. In nur zwei Fällen werden positive d' berechnet[9].

Grundsätzlich nehmen die in den Unterabbildungen 4.15a bis 4.15d dargestellten d' mit einem zunehmenden Anteil an fehlenden Werten ab. Alle Kurven beginnen bei einem leicht negativen d' für 5% fehlende Werte und sinken von dort ab. Eine Erhöhung der Merkmalsanzahl verursacht zwei Veränderungen. Zum einen werden die berechneten Effektgrößen betragsmäßig kleiner. Zum anderen findet eine Spreizung der dargestellten Kurven statt. Dies resultiert darin, dass 42 der 120 betrachteten Fälle keine bedeutsamen Unterschiede zwischen einer Hot-Deck-Imputation mit und ohne Donor-Limit aufweisen. Eine Reduzierung der Objektanzahl führt im Allgemeinen zu einer kleinen betragsmäßigen Reduzierung der Effektstärken.

[9] Diese sind jedoch innerhalb der Trivialitätsgrenzen.



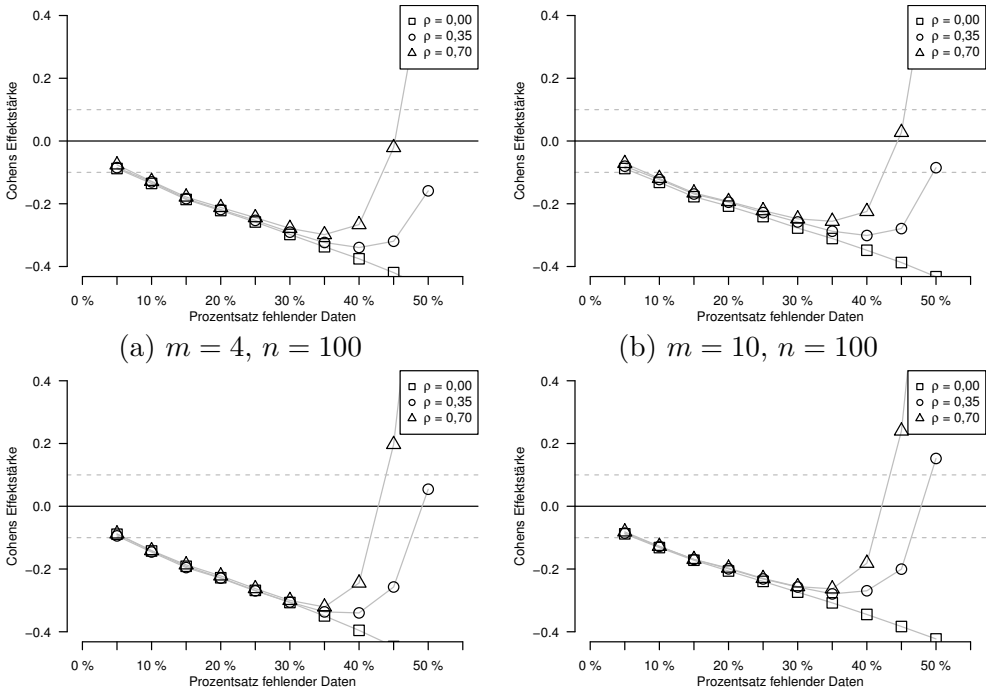
(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.15: Auswirkung von *MCAR*-Ausfall auf den Kontingenzkoeffizienten bei nominal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

MAR 1:2-Ausfall

Der Effekt des Donor-Limits auf die Ausprägungshäufigkeitsschätzung bei einem nominalen Merkmal, das dem verschärften Ausfallmechanismus *MAR 1:2* ausgesetzt wurde, ist weder durchweg vorteilhaft noch durchweg von Nachteil. Dies ist daran zu erkennen, dass nicht alle berechneten d' in Abbildung 4.16 dasselbe Vorzeichen aufweisen. Dennoch ist die Anzahl der Fälle, bei dem ein Donor-Limit zu wesentlichen Nachteilen in der Imputationsqualität führt, mit 7 Fällen klein. Bedeutsame Vorteile bietet das Verfahren mit Donor-Limit in 97 Fällen, insbesondere, wenn der Anteil fehlender Werte nicht größer als 40% ist. Fehlen mehr als 40% der Werte in dem zu imputierenden Merkmal, kommt es darauf an, wie stark die Korrelation innerhalb der simulierten Datenmatrix ist. Für $\rho = 0{,}00$ ist ein Donor-Limit nie nachteilig. Bei den anderen beiden simulierten Korrelationen ist ab ca. 35% fehlender Werte eine Trendwende zu beobachten. Hier ist dann die Rate an Verlust von Vorteilhaftigkeit für ein größeres ρ höher.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

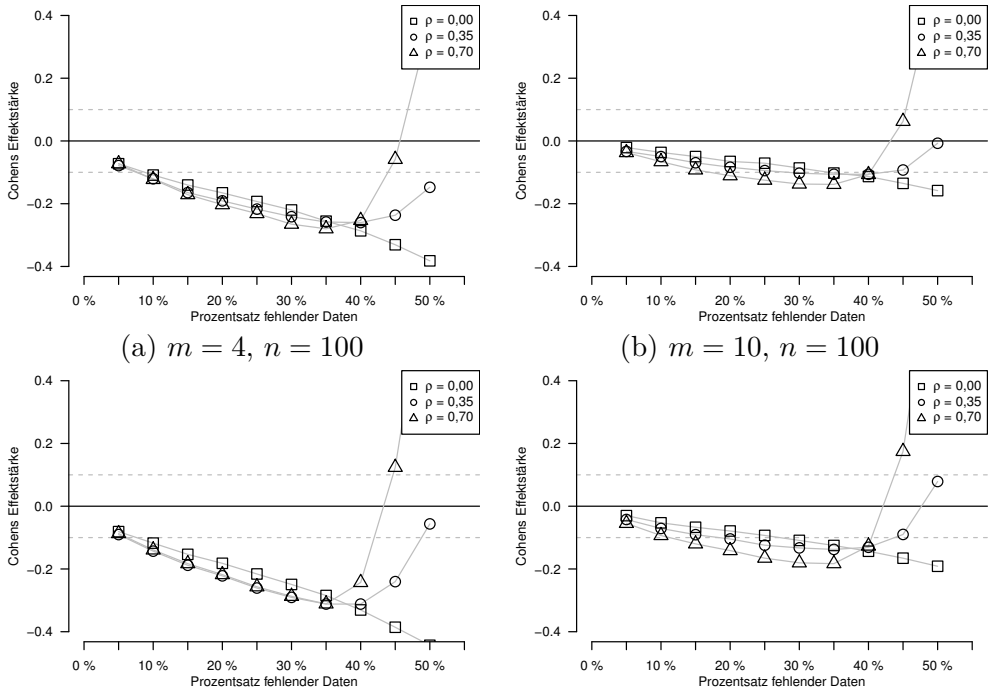
Abbildung 4.16: Auswirkung von *MAR 1:2*-Ausfall auf die Ausprägungshäufigkeit eines nominal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

Der Einfluss einer Variierung der Merkmals- oder Objektanzahl ist grundsätzlich klein. Eine Erhöhung der Merkmalsanzahl führt grundsätzlich zu einer Erhöhung der d'. Eine Erhöhung der Objektanzahl führt meist zu einer Verringerung der Effektstärken, ausgenommen, wenn größere Anteile an Werten fehlen. Hier ist eine kleine Spreizung der in Abbildung 4.16 dargestellten Kurven zu erkennen.

Grundsätzlich ergeben sich im direkten Vergleich zu der Ausprägungshäufigkeitsschätzung bei dem *MCAR*-Datenausfall nur geringe Unterschiede. Für die meisten Fälle ist bei dem *MAR 1:2* ein Donor-Limit deutlich mehr von Vorteil als beim *MCAR*-Datenausfall. Lediglich bei einer extremen Anzahl an fehlenden Werten und dem $\rho = 0{,}70$ sind zudem im Verlauf der d' qualitative Unterschiede vorhanden. So ist hier, beim intensiveren Ausfallmechanismus, bei neun Kovariaten bereits ab 45% das Verfahren ohne Donor-Limit bedeutsam besser.

Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits auf die Schätzung des normierten Kontingenzkoeffizenten zwischen dem imputierten und einem weiteren nominalen Merkmal ist im Falle des intensiver-

en *MAR 1:2*-Ausfallmechanismus nicht eindeutig. Zwar sind die berechneten d' – wie in Abbildung 4.17 zu erkennen – meist negativ[10], jedoch treten auch in den Extremfällen deutlich positive d' auf[11].



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.17: Auswirkung von *MAR 1:2*-Ausfall auf den Kontingenzkoeffizienten bei nominal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Wie auch schon bei dem *MCAR*-Ausfallmechanismus führt insbesondere eine größere Anzahl an Merkmalen zu trivialen Unterschieden zwischen der Imputation mit und ohne Donor-Limit[12]. Ist nur eine kleinere Anzahl an Kovariaten vorhanden, so ist der Vorteil einer Hot-Deck-Imputation mit Donor-Limit größer.

Im direkten Vergleich zum *MCAR*-Ausfallmechanismus kann festgehalten werden, dass die Unterschiede im Allgemeinen klein sind. Insbesondere treten diese bei den größeren Anteilen an fehlenden Werten auf. Vorwiegend ab 35% fehlender Werte äußert sich die Intensivierung des Ausfallmechanismus durch eine größere Spreizung der Kurven für die unterschiedlichen ρ. Im Vergleich

[10] 112 der 120 berechneten Effektstärken sind negativ.

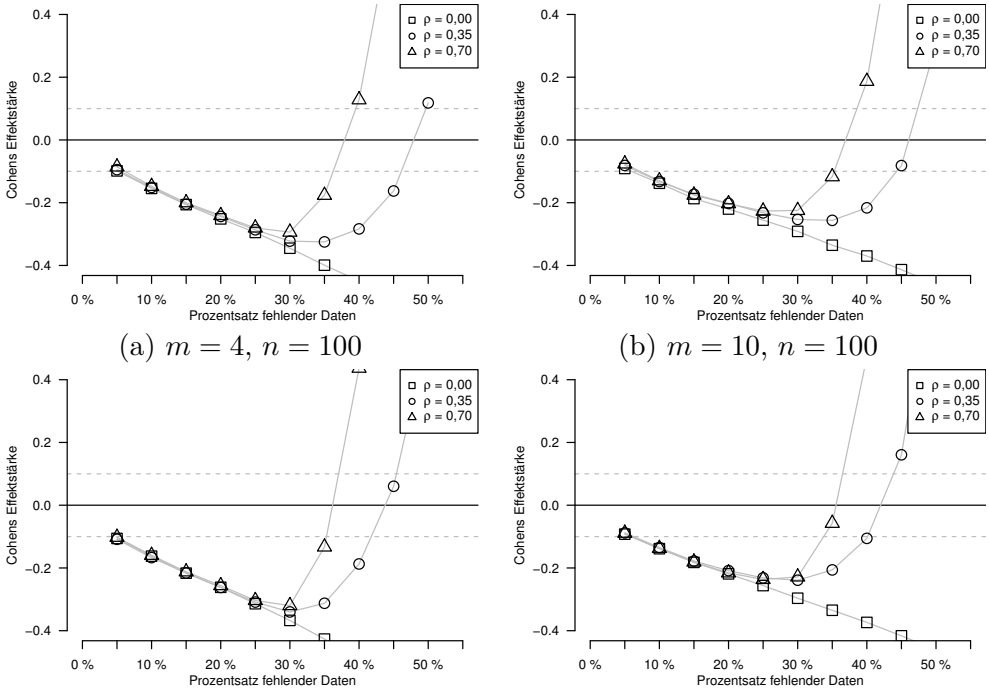
[11] 6 der 120 berechneten Effektstärken sind größer als 0,1.

[12] Insgesamt weisen 37 der betrachteten Fälle d' von trivialer Größe auf.

der Abbildungen 4.15 und 4.17 fällt zudem insbesondere der Vorzeichenwechsel für $\rho = 0{,}70$ und eine hohe Anzahl an fehlenden Werten auf.

MAR 1:4-Ausfall

Werden jene simulierten Fälle betrachtet, bei denen eine Schätzung der Häufigkeit einer Ausprägung bei einem nominalen Merkmal unter dem intensivsten Ausfallmechanismus stattfindet, so ist eine Hot-Deck-Imputation mit einem Donor-Limit zwar meistens – aber nicht immer – die beste Wahl. Wiederum zeigt sich bei diesem Ausfallmechanismus, dass insbesondere in wenigen der extremen Fälle eine Hot-Deck-Imputation ohne Donor-Limit von Vorteil ist. Selbst bei dem intensivsten Ausfallmechanismus ist bis 35% fehlender Werte immer eine Imputation mit Donor-Limit vorteilhaft.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.18: Auswirkung von *MAR 1:4*-Ausfall auf die Ausprägungshäufigkeit eines nominal skalierten Merkmals; Effektstärken vs. Anteil fehlender Werte

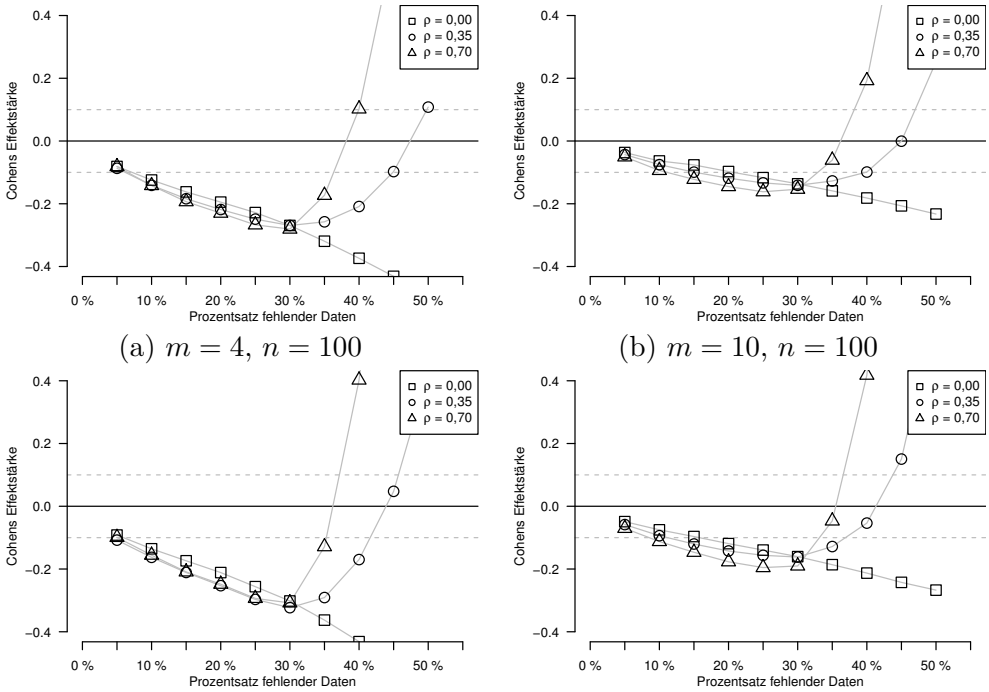
Der Kurvenverlauf ist bei allen Unterabbildungen 4.18a bis 4.18d sehr ähnlich. Alle Kurven beginnen immer bei $d' \approx -0{,}1$ und fallen zunächst mit einem zunehmenden Anteil fehlender Werte weiter ab. Während dies sich für $\rho = 0{,}00$ bis zu einem Anteil fehlender Werte von 50% fortsetzt, wird für $\rho = 0{,}35$ und

$\rho = 0{,}70$ bei einem Anteil fehlender Werte von ca. 0,3 ein Minimum erreicht. Die diesem Minimum zugehörigen d' betragen ungefähr $-0{,}3$. Nach Erreichen dieser Minima steigen die entsprechenden Kurven wieder an. Der Anstieg richtet sich nach dem simulierten n, m und ρ. Grundsätzlich gilt, dass eine Erhöhung aller drei Parameter zu einer Erhöhung der Anstiege in den d' führt. Bei $\rho = 0{,}00$ führt eine Erhöhung von n zu einer weiteren Verringerung des Anstiegs, so dass eine Erhöhung von n eine Spreizung der Kurven in Abhängigkeit von ρ verursacht.

Werden die Ausprägungshäufigkeitenschätzungen über die betrachteten Ausfallmechanismen hinweg verglichen, so ist grundsätzlich eine Intensivierung der Effektstärken festzustellen. Ist für eine bestimmte Faktorstufenkombination bei dem *MCAR*-Ausfallmechanismus ein Donor-Limit von Vorteil, so ist dies grundsätzlich unter dem *MAR 1:4*-Ausfallmechanismus deutlich besser. Die Ausnahme hierfür stellen einige der betrachteten Fälle dar, wenn die Korrelation innerhalb der simulierten Datenmatrix $\rho = 0{,}35$ oder $\rho = 0{,}70$ beträgt. Hier ist – abweichend von der vorherigen Aussage – festzustellen, dass insbesondere der Anstieg nach dem Minimum durch eine Intensivierung des Ausfallmechanismus erhöht wird. So treten bei dem *MAR 1:4*-Ausfallmechanismus in Randfällen d' auf, die andeuten, dass eine Hot-Deck-Imputation ohne Donor-Limit für die Parameterschätzung von Vorteil ist.

Die Auswirkung der variierten Faktoren auf die Vorteilhaftigkeit des Donor-Limits bei der Schätzung des normierten Kontingenzkoeffizenten zwischen dem imputierten und einem weiteren nominalen Merkmal ist auch im Falle des stärksten Ausfallmechanismus, *MAR 1:4*, nicht eindeutig. Die berechneten d' sind insbesondere in Abhängigkeit der vorhandenen Korrelation innerhalb der simulierten Matrizen entweder positiv oder negativ. In 76 von 120 Fällen ist die Imputation mit Donor-Limit überlegen, in 17 Fällen ist eine Imputation ohne Donor-Limit besser und in den verbleibenden 27 Fällen sind die Effektstärken betragsmäßig trivial. Somit ist auch beim intensivsten Ausfallmechanismus meist ein Donor-Limit für die Imputationsqualität nicht von Nachteil.

Grundsätzlich ist der Kurvenverlauf in den Unterabbildungen 4.19a bis 4.19d für gleichbleibende ρ ähnlich. Wenn $\rho = 0{,}00$, sind alle d' stets negativ. Der Kurvenverlauf beginnt innerhalb der Trivialitätsgrenzen und fällt für einen zunehmenden Anteil fehlender Werte weiter ab.



(a) $m = 4$, $n = 100$ (b) $m = 10$, $n = 100$

(c) $m = 4$, $n = 200$ (d) $m = 10$, $n = 200$

Abbildung 4.19: Auswirkung von *MAR 1:4*-Ausfall auf den Kontingenzkoeffizienten bei nominal skalierten Merkmalen; Effektstärken vs. Anteil fehlender Werte

Hot-Deck-Imputation ohne Donor-Limit ist meist wieder unter den extremeren der betrachteten Fälle von Vorteil. So entstehen $d' \geq 0{,}1$ insbesondere bei den größeren ρ und einem größeren Anteil fehlender Werte. Diese Effekte und der damit assoziierte Vorteil bei der Imputationsqualität werden, in Abhängigkeit des Datenausfalls, zu groß, um in Abbildung 4.19 dargestellt zu werden. Beispielsweise entspricht das d' für ein $\rho = 0{,}70$ und einer 200×10 simulierten Matrix bei einem Datenausfall von 50% einem Wert[13] von ca. 2,78.

Im Allgemeinen führt eine Erhöhung der Merkmalsanzahl zu einer Erhöhung der Effektstärken, erkennbar an einer Vertikalverschiebung der Kurven in Abbildung 4.19 nach oben. Die Auswirkungen einer Erhöhung der Objektanzahl sind bei einer geringeren Merkmalsanzahl besser zu erkennen als bei einer größeren. Beim Vorhandensein von mehr Objekten ist auch eher ein Donor-Limit von Vorteil. Dies ist am deutlichsten im Vergleich der Kurven für $\rho = 0{,}00$ beziehungsweise $\rho = 0{,}70$ in den Unterabbildungen 4.19a und 4.19c

[13] Dies ist das größte d', welches für die Schätzung des normierten Kontingenzkoeffizienten zwischen dem imputierten und einem weiteren nominalen Merkmal berechnet wurde.

zu erkennen. Insgesamt ist die Auswirkung eher klein und wird erheblich von der Auswirkung einer Merkmalserhöhung überlagert.

Auch bei der Korrelationsschätzung ergeben sich im Allgemeinen und im direkten Vergleich zu dem *MAR 1:2*-Ausfallmechanismus geringere Unterschiede als zum *MCAR*-Ausfallmechanismus. In den simulierten Fällen bewirkt der Übergang von *MCAR* zu *MAR 1:2* zu *MAR 1:4* tendenziell eine Verstärkung der vorhandenen Effekte auf die berechneten d'. Nur in seltenen Fällen wird ein Vorzeichenwechsel durch die Variation des Ausfallmechanismus herbeigeführt. Diese seltenen Fälle sind zudem durch die extremen Ausprägungen anderer Faktoren charakterisiert[14].

4.2.2.4 Auswirkungen der Skalenvariierung

Um die Wirkung einer Skalenvariierung auf die Vorteilhaftigkeit des Donor-Limits zu beurteilen, werden die berechneten d' für die univariaten Verteilungsparameter und multivariaten Zusammenhangsmaße getrennt betrachtet. Die Darstellung nimmt Bezug auf die in den Abschnitten 4.2.2.1, 4.2.2.2 und 4.2.2.3 identifizierten Effekte.

Univariate Verteilungsparameter

Ob eine Hot-Deck-Imputation mit oder ohne Donor-Limit eher von Vorteil ist, kann für den Fall, dass die betrachteten univariaten Verteilungsparameter von übergeordnetem Interesse sind, eindeutig beantwortet werden. Betrachtet man jene in Tabelle 4.6 dargestellten Häufigkeiten, ist ersichtlich, dass eine Imputation mit Donor-Limit am häufigsten von Vorteil ist. Zudem kann dieser Tabelle entnommen werden, dass für ein nominales Merkmal eine Hot-Deck-Imputation ohne Donor-Limit häufiger in Frage kommt als für ein ordinales oder metrisches Merkmal. Die vier Fälle, für die eine Imputation ohne Donor-Limit von Vorteil ist, sind für Varianz und Quartilsabstand dieselben. Sie stellen auch im Falle dessen, dass die Häufigkeit einer bestimmten Ausprägung bei einem nominalen Merkmal geschätzt werden soll, Situationen dar, bei denen eine Imputation ohne Donor-Limit überlegen ist[15]. Da diese vier Fälle sowie die weiteren 20 Konstellationen, unter denen der nominale Parameter

[14] Beispielsweise, wenn 50% der Werte fehlen.

[15] Die vier Konstellationen können den Abbildungen 4.6, 4.12 und 4.18 entnommen werden.

besser bei einer Hot-Deck-Imputation ohne Donor-Limit geschätzt wird, alle Extremfälle darstellen[16], kann darauf geschlossen werden, dass eine Skalendegression einen geringen Einfluss auf die Vorteilhaftigkeit des Donor-Limits ausübt. Diese Aussage kann durch eine Betrachtung der mittleren d' für die unterschiedlichen Parameter unterstützt werden. Für Varianz, Quartilsabstand und Ausprägungshäufigkeit betragen diese $-0{,}209$; $-0{,}233$ und $-0{,}221$.

	$d' \in$		
	$(-\infty;\ -0{,}1]$	$(-0{,}1;\ 0{,}1)$	$[0{,}1;\ \infty)$
Varianz	307	49	4
Quartilsabstand	316	40	4
Ausprägungshäufigkeit	292	44	24

Tabelle 4.6: Häufigkeiten von bestimmten Effektstärken für univariate Verteilungsparameter

Multivariate Zusammenhangsmaße

Bezüglich dessen, ob eine Hot-Deck-Imputation mit Donor-Limit einer ohne Donor-Limit überlegen ist, kann, falls die Schätzung mulitvariater Zusammenhangsmaße die Imputationsqualität bestimmt, nicht so eindeutig wie im vorherigen Fall beantwortet werden. Dies folgt unmittelbar aus der Tabelle 4.7. Zwar ist meist eine Imputation mittels Donor-Limit nicht von wesentlichem Nachteil oder sogar vorteilhaft, jedoch ist die Menge an Fällen, bei denen eine Imputation mit keinem Donor-Limit überlegen ist, größer als bei den univariaten Parametern. Durch eine Skalenprogression ergeben sich gewisse Verschiebungen in den Häufigkeiten. Ist das Merkmal metrisch skaliert, so ist es häufiger von Vorteil, kein Donor-Limit zu verwenden als bei einer ordinalen oder nominalen Merkmalsskalierung. Die Fälle, in denen die Imputation ohne Donor-Limit von Vorteil ist, stellen im gewissen Maße wiederum Extremfälle dar[17].

Durch eine Analyse der Abbildungen aus den Abschnitten 4.2.2.1, 4.2.2.2 und 4.2.2.3 kann zudem darauf geschlossen werden, dass eine Skalendegression

[16] Alle Fälle haben gemein, dass sie bei einem der beiden *MAR*-Ausfallmechanismen mit 40% oder mehr fehlenden Werten auftreten.

[17] Die 88 Fälle aus Tabelle 4.7 treten bei mehr als 30% Ausfall und präferiert bei hohem ρ auf.

einen geringen Einfluss auf die Vorteilhaftigkeit des Donor-Limits ausübt. Diese Aussage kann durch eine Betrachtung der mittleren d' für die unterschiedlichen Parameter unterstützt werden. Für Varianz, Quartilsabstand und Ausprägungshäufigkeit betragen diese entsprechend $-0{,}216$; $-0{,}237$ und $-0{,}156$.

	$d' \in$		
	$(-\infty;\ -0{,}1]$	$(-0{,}1;\ 0{,}1)$	$[0{,}1;\ \infty)$
Korrelation	155	165	40
Rangkorrelation	260	75	25
Kontingenzkoeffizent	231	106	23

Tabelle 4.7: Häufigkeiten von bestimmten Effektstärken für multivariate Zusammenhangsmaße

4.2.3 Zusammenfassung

Die im Rahmen dieses Abschnitts durchgeführte konfirmatorische Studie bestätigt, dass die Nutzung eines Donor-Limits nicht immer in eine höhere Imputationsqualität mündet. Vielmehr zeigte sich, wie auch in der Voruntersuchung, dass eine Hot-Deck-Imputation ohne Spenderbeschränkung zu besseren Imputationswerten führen kann.

Wie auch bei der Voruntersuchung konnten aufgrund der detaillierten Betrachtung der Ergebnisse dieser Simulationsstudie allgemeingültige Aussagen formuliert werden. Faktoren, die primär beeinflussen, ob ein Donor-Limit von Vorteil ist, sind der Ausfallmechanismus, der Anteil fehlender Werte, die Korrelationsstruktur der Datenmatrix und ob eher die Schätzung von univariaten Parametern oder multivariate Zusammenhangsmaße von Interesse sind. Die Skalierung des Merkmals mit fehlenden Werten, die Anzahl der Objekte und die Anzahl der Merkmale spielen hingegen eine untergeordnete Rolle. Kennzeichnend für die Ergebnisse ist zudem, dass insbesondere die Wechselwirkungen zwischen den Faktoren und nicht deren Haupteffekte dominieren.

Die detaillierte Darstellung der Ergebnisse zeigte jene bei der Voruntersuchung vermuteten, interessanten Wirkungen auf und bestätigte abermals die vorher ermittelten Haupteffekte. Zusammenfassend ergeben sich aus den Ergebnissen des Abschnitts 4.2.2 insbesondere folgende Aussagen über die Kon-

stellationen, bei denen ein Donor-Limit vorteilhaft ist:

1. Für weniger als 35% fehlender Daten ist das Verfahren mit Donor-Limit nie erheblich schlechter ($d' < 0{,}1$) als das Verfahren ohne Donor-Limit.

2. Fehlen weniger als 25% der Werte des zu imputierenden Merkmals, so ist eine Hot-Deck-Imputation mit Donor-Limit immer überlegen ($d' < 0$).

3. Ist die Datenstruktur schwach ($\rho \approx 0{,}00$), so ist eine Hot-Deck-Imputation mit Donor-Limit immer überlegen ($d' < 0$).

4. Ist eine Imputation mit Donor-Limit vorteilhaft, so steigt tendenziell dieser Vorteil mit dem Anteil fehlender Werte.

5. Mit steigender Anzahl an Kovariaten werden die Effektstärken kleiner.

6. Je weniger Objekte vorhanden sind, um so kleiner sind die Effektstärken.

7. Je extremer der vorliegende Fall, gemessen an Ausfallmechanismus, Anteil fehlender Werte und Datenstruktur, ist, um so eher ist eine Imputation ohne Donor-Limit von Vorteil.

8. Ist eine Imputation ohne Donor-Limit von merklichem Vorteil ($d' \geq 0{,}1$), so ist dieser Vorteil tendenziell um einige Größenordnungen größer als jene Vorteile, die mit einem Donor-Limit erzielbar sind[18].

9. Bei der Entscheidung, ob ein Donor-Limit verwendet werden soll, spielt die Frage nach der Skalierung des zu imputierenden Merkmals eine untergeordnete Rolle.

10. Die Schätzung eines univariaten Parameters profitiert häufiger und stärker von einem Donor-Limit als die Schätzung eines multivariaten Zusammenhangsmaßes.

[18] Das kleinste d' beträgt $-0{,}643$, während das größte bei 3,825 liegt.

Kapitel 5

Schlussbemerkungen

Dieses letzte Kapitel von „Hot-Deck-Verfahren zur Imputation fehlender Daten – Auswirkungen des Donor-Limits“ widmet sich abschließenden Worten für die Arbeit. Diesbezüglich beinhaltet Abschnitt 5.1 eine Zusammenfassung. Im Abschnitt 5.2 werden die Erkenntnisse der empirischen Untersuchungen kritisch gewürdigt. Die Arbeit schließt im Abschnitt 5.3 mit einem Ausblick auf den weiteren Forschungsbedarf ab.

5.1 Zusammenfassung

Das Ziel dieser Arbeit bestand darin zu ergründen, ob und unter welchen Bedingungen die Verwendung eines Donor-Limits Vorteile bietet. Dies erforderte zunächst eine Auseinandersetzung mit den Grundlagen fehlender Daten im Allgemeinen und mit Hot-Deck-Verfahren im Speziellen. Nach einer Erarbeitung dieser und dem Aufzeigen des Forschungsbedarfs wurden die Auswirkungen des Donor-Limits empirisch untersucht.

Die Auseinandersetzung mit den Grundlagen fehlender Daten im Allgemeinen erfolgte anhand der vier zentralen Fragen der Missing-Data-Problematik im Kapitel 2. Die Frage: „Weshalb fehlen die Daten?“ wurde in Abschnitt 2.1 durch die Ausfallursachen beantwortet. Hier konnte gezeigt werden, dass sich zur besseren Erkennung der Kausalzusammenhänge, welche in einen Datenausfall münden, die Datenerhebung und -auswertung als soziotechnisches System (in Anlehnung an Trist und Bamford, 1951) modellieren lässt. Den Kern dieses Systems bildet die Kommunikationsstruktur, welche wiederum mittels des

Shannon-Weaver-Modells (Shannon und Weaver, 1949, S. 34) darstellbar ist. Zusammen mit dem Untersuchungsdesign, welches extern auf das Datenerhebungssystem Einfluss nimmt, konnten sämtliche Ausfallursachen systematisiert werden. Hit Hilfe des Konzepts der Muster fehlender Daten wurde die Frage „Wo fehlen die Daten?“ beantwortet. Hierzu wurden sechs Ausfallmuster vorgestellt, die typischerweise in der Literatur diskutiert werden. Die Antwort auf die Frage „Wie fehlen die Daten?“ erfolgte anhand der von Rubin (1976a) aufgestellten Theorie der Ausfallmechanismen. Demnach werden Ausfallursachen als stochastische Prozesse betrachtet, deren Eigenschaften maßgeblich für den weiteren Umgang mit dem Datenausfall sind. An dieser Stelle konnten Sonderfälle erarbeitet werden, die insbesondere im Falle von Querschnittsstudien die möglichen Wirkungsbeziehungen zwischen A^{obs} und A^{mis} besser abbilden. Des Weiteren wurde mit den Gedankenexperimenten aus Abschnitt 2.3.4 ein Beitrag zur Diskussion von Enders (2010, S. 14–17) sowie Schafer und Graham (2002, S. 173) über die Plausibilität des NMAR-Ausfallmechanismus geleistet. Abgeschlossen wurde das Kapitel 2 mit Erläuterungen zum möglichen Umgang mit fehlenden Daten. Hier wurden unterschiedliche Systematisierungsansätze für Missing-Data-Methoden sowie unterschiedliche Eliminierungs- und Imputationsverfahren vorgestellt.

Kapitel 3 begann mit einer Darstellung der historischen Entwicklung und Herkunft von Hot-Deck-Verfahren. Es folgte die Erarbeitung einer Definition für Hot-Deck-Verfahren mittels einer umfassenden Sichtung relevanter Literatur und einer Abwägung der verschiedensten Verständnisse, die in dieser vorherrschen. Gemäß dieser Definition besteht jeder Hot-Deck-Algorithmus aus drei Komponenten:

1. Spender- beziehungsweise Empfänger-Identifikation
2. Spender-Empfänger-Zuordnung
3. Verdoppelung der Spenderwerte.

Ferner konnte aus der Literatur extrahiert werden, dass sich Hot-Deck-Verfahren lediglich in den vier folgend aufgeführten zentralen Eigenschaften unterscheiden:

1. Definition von Ähnlichkeit
2. Stochastizität
3. Behandlung mehrerer Merkmale
4. Mehrfachverwendung der Spender.

Zur Darstellung, welche konkreten Ausprägungen diese vier Eigenschaften haben können, wurden in Abschnitt 3.2 die in der Literatur existenten Möglichkeiten präsentiert. Neben einem eigenen Vorschlag, wie bei der Distanzberechnung zwischen Spendern und Empfängern den Forderungen von Ford (1983, S. 186) nachgekommen werden kann, mündete die Darstellung der Eigenschaften in die Entwicklung eines Optimierungsproblems. Diese neue Verfahrensvariation, welche sich durch die Darstellung von Hot-Deck-Methoden als ganzzahliges Optimierungsproblem ergibt, wurde im Abschnitt 3.3 vorgestellt. Durch eine Simulationsstudie konnte gezeigt werden, dass eine Lösung dieses Optimierungsproblems insbesondere für die Schätzung multivariater Parameter von Vorteil sein wird.

Nach der Erarbeitung notwendiger Grundlagen in den Kapiteln 2 und 3 widmete sich Kapitel 4 mit Hilfe von zwei breit aufgestellten Simulationsstudien der empirischen Untersuchung des Donor-Limits und dessen Auswirkungen. So konnten bereits in der Voruntersuchung (Abschnitt 4.1) anhand der Haupteffekte von vermuteten Einflussfaktoren gewisse Aussagen getroffen werden. Die Ergebnisse deuten unmissverständlich darauf hin, dass Situationen existieren, unter denen der Einsatz des Donor-Limits sinnvoll ist. Als einflussreiche Faktoren konnten der Anteil fehlender Werte, die Anzahl der Imputationsklassen, die Parameter, die es zu schätzen gilt, und das verwendete Hot-Deck-Verfahren identifiziert werden. Aufgrund der vorhandenen Wechselwirkungen wurde für die Entwicklung konkreter Empfehlungen eine nachgelagerte konfirmatorische Studie durchgeführt. Diese tiefgehende Untersuchung erfolgte im Abschnitt 4.2. Da in Simulationstudien, die eine hinreichende Anzahl an wiederholten Berechnungen aufweisen, durch eine nochmalige Durchführung keine veränderten Ergebnisse zu erwarten sind, wurde die konfirmatorische Studie – basierend auf den Ergebnissen und Erkenntnissen der Voruntersuchung – unter Einbezug weiterer Faktoren gestaltet. Um in dieser Studie nicht nur die Haupteffekte,

sondern alle Wechselwirkungen zu erfassen, erfolgte die Auswertung der Ergebnisse unter Betrachtung aller 2.160 möglichen Faktorstufenkombinationen. Insgesamt lassen sich die wesentlichen Ergebnisse der Voruntersuchung – gemeinsam mit denen der konfirmatorischen Studie – im Folgenden thesenförmig festhalten:

1. Ob eine Hot-Deck-Imputation mit oder ohne Donor-Limit durchgeführt wird, beeinflusst die Imputationsergebnisse.

2. Es kann erwartet werden, dass ein Donor-Limit die Imputationsvarianz minimiert.

3. Die Ergebnisse eines Hot-Deck-Verfahrens, das einen Spender zufällig auswählt, wird durch ein Donor-Limit nicht beeinflusst.

4. Ein Donor-Limit wirkt sich weder auf die Schätzung des Mittelwerts eines kardinalen Merkmals noch auf die Schätzung des Medians eines ordinalen Merkmals aus.

5. Ein Donor-Limit verbessert grundsätzlich die Schätzung von Variabilitäts- und Zusammenhangsmaßen.

6. Die Effekte einer reduzierten Stichprobengröße wirken sich gleichmäßig auf die Imputation mit und ohne Donor-Limit aus.

7. Bei einer geringen und moderaten Anzahl an fehlenden Werten ist von einer Imputation mit Donor-Limit nie abzuraten.

8. Ob ein Merkmal nominal, ordinal oder kardinal skaliert ist, hat eine vernachlässigbar kleine Auswirkung auf die Vorteilhaftigkeit des Donor-Limits.

9. Je mehr Kovariaten, mit deren Hilfe die Ähnlichkeiten bestimmt werden, vorhanden sind, umso kleiner fallen die Vor- und Nachteile des Donor-Limits aus.

10. Umso stärker die Struktur der Daten ist, desto eher ist ein Donor-Limit bei einem großen Anteil fehlender Daten von Nachteil.

Zusammenfassend lässt sich festhalten, dass folgender Grundsatz für die Verwendung des Donor-Limits gilt:

> *Je extremer die durch den Datenausfall bedingte Situation ist, desto eher sollte auf ein Donor-Limit verzichtet werden. Jedoch ist eine Hot-Deck-Imputation mit Donor-Limit meist sinnvoll und bei weniger als 25% fehlender Werte stets zu empfehlen.*

5.2 Kritische Würdigung

Die in dieser Arbeit gewonnenen Erkenntnisse über das Donor-Limit bei der Hot-Deck-Imputation wurden mittels Simulation, in deren Rahmen Daten und Ausfallmechanismen künstlich erzeugt wurden, erzielt. Während die Simulation synthetischer Ausfallmechanismen unumgänglich ist, da die wahren Werte bekannt sein müssen, um die Imputationsqualität zu bestimmen, existiert bei den Daten mindestens eine weitere Option. Anstatt synthetische Datensätze zu generieren, hätten einzelne reale Datensätze verwendet werden können. Diese Vorgehensweise, wenngleich hier reale und keine realistischen Strukturen verwendet werden, würde jedoch als Art Fallstudie weniger generalisierbare Ergebnisse liefern. Gewisse Zusammenhänge und Eigenschaften der Daten müssten hingenommen werden und der Einfluss bestimmter Parameter könnte, im Gegensatz zum gewählten Ansatz, nie so genau bestimmt werden.

Des Weiteren ist eine Simulation immer durch ihre Annahmen begrenzt. Hier wurde bei der Datenverteilung und dem Ausfallmechanismus Wert auf realitätsnahe Fälle gelegt. Dennoch sind weitere Verteilungen und Ausfallmechanismen möglich, deren Wahl nicht weniger angemessen wäre.

Ferner ist diese Untersuchung durch die Wahl der Parameter und deren Variationsstufen begrenzt. Zwar genügt für die Generierung von Tendenzaussagen die Variation der gewählten Parameter auf der jeweils gewählten Anzahl von Stufen; dennoch sind definitive Aussagen stets nur für die über 2.160 betrachteten Fälle möglich. Ähnlich ist es mit den gewählten Parametern. Es wurden über die Voruntersuchung und die konfirmatorische Studie hinweg zwar zehn Faktoren betrachtet, trotzdem wären hier weitere möglich.

5.3 Ausblick

Nicht nur die kritische Würdigung, sondern auch die in den Kapiteln 2 und 3 geleistete Literaturarbeit zeigt den möglichen weiteren Forschungsbedarf auf.

Diese Arbeit stellte sich zur Aufgabe zu untersuchen, ob und unter welchen Bedingungen die Verwendung eines Donor-Limits Vorteile bietet. Hinsichtlich dieser Zielstellung ist diese Arbeit einzigartig, nicht nur, weil bis dato die Wirkungen eines Donor-Limits noch nicht untersucht wurden, sondern auch, weil die unterschiedlichen Möglichkeiten Hot-Deck-Imputation durchzuführen, grundsätzlich noch nicht systematisch betrachtet wurden. Bedenkt man, dass bereits ohne Berücksichtigung der konkreten Klassen- und Distanzberechnungsmethoden fast 100 Varianten Hot-Deck-Verfahren durchzuführen, existieren, ist dies auch nicht verwunderlich. So ist es denkbar, dass zukünftige Arbeiten die verbleibenden drei Eigenschaften der Hot-Deck-Verfahren betrachten und ergründen, wie jeweils am besten zu verfahren ist. Beispielsweise könnten die verschiedenen Möglichkeiten, die Spender-Empfänger-Ähnlichkeiten zu bestimmen, untersucht werden. Auch eine Betrachtung dessen, unter welchen Bedingungen eine sequentielle oder simultane Behandlung der Merkmale mit fehlenden Werten überlegen ist, könnte interessante Ergebnisse liefern. Ausstehend sind jedoch noch grundsätzlichere Untersuchungen, etwa wie sich die Imputationsqualität von Hot-Deck-Verfahren mit der von anderen Methoden vergleichen lässt.

Offen bleiben auch praktische Tätigkeiten. Da eine Analyse von Daten im Allgemeinen und hierdurch die Behandlung fehlender Daten im Speziellen ohne den Einsatz von Software heute nicht mehr denkbar ist, ist die Verfügbarkeit der vorgestellten Algorithmen für ihre Anwendung von entscheidender Bedeutung. Sämtliche der in der Literatur vorgeschlagenen Hot-Deck-Verfahren wurden nie allgemein zugänglich gemacht. Insbesondere ältere Algorithmen mit historischer Bedeutung sind nicht als Software verfügbar. Um die Reproduzierbarkeit der Forschung zu gewährleisten, wurden die hier verwendeten Imputationsalgorithmen in einem frei verfügbaren R-Paket **HotDeckImputation** (Joenssen, 2013) veröffentlicht. Eine Implementierung weiterer Hot-Deck-Verfahren in dieses oder einem anderen Softwarepaket wäre sicher auch für andere Forscher lohnenswert.

Anhang A

Stichprobe der 2009 ACS

Erläuterungen: Die in der Tabelle A.1 dargestellte Datenmatrix gibt eine Stichprobe elf ausgewählter Merkmale der American Community Survey (ACS) aus 2009 vom 28. September 2010 (Ruggles et al., 2010; U.S. Bureau of the Census, 2010) wieder. Die ACS ist eine jährliche Befragung von einem Prozent der U.S.-Amerikanischen Bevölkerung durch das U.S. Census Bureau. Zur Imputation fehlender Werte, die primär durch Antwortverweigerung der Befragten oder Löschung inkonsistenter Werte verursacht werden, wird eine Form der Nearest-Neighbor-Hot-Deck-Imputation innerhalb von Schichten verwendet. Merkmale, die fehlende Werte aufweisen, werden sequentiell betrachtet (U.S. Bureau of the Census, 2010). Ergebnisse aus dieser Befragung werden zur Planung öffentlicher Haushalte verwendet und der Öffentlichkeit zur Verfügung gestellt. Die Werte der Merkmale definieren sich wie folgt:

- **PERWT:** Personen Gewicht; kardinal
- **REGION:** Zensus Region; nominal polytom mit S = Süden; N = Norden; W = Westen; O = Osten; C = Zentrum
- **SEX:** Geschlecht; nominal dichotom mit M = Männlich; F = Weiblich
- **AGE:** Alter; kardinal
- **EDUCD:** Bildungsniveau (in Jahren); kardinal
- **LANGUAGE:** Muttersprache; nominal polytom mit Eng = Englisch; Sp = Spanisch; Fr = Französisch; Ger = Deutsch

- **TRANTIME:** Arbeitsweg (in Minuten); kardinal
- **UHRSWORK:** Wochenarbeitszeit (in Stunden); kardinal
- **INCTOT:** Individualeinkommen (in USD); kardinal
- **FTOTINC:** Familieneinkommen (in USD); kardinal
- **INCSS:** Erhaltene Staatliche Transferleistung (Social Security; in USD); kardinal

Damit diese Datenmatrix zur Veranschaulichung der in dieser Arbeit präsentierten Algorithmen dienen kann, wird unter den vier folgend beschriebenen Situationen unterschieden. Hierbei wurde berücksichtigt, dass mindestens 50% der Objekte keine fehlenden Werte aufweisen ($p \leq 12$).

- **Fall 1:**
  Die Daten fehlen MCAR. Bei den Merkmalen SEX und LANGUAGE fehlt eine Ausprägung mit einer Einzelwahrscheinlichkeit von 0,2.

$$Pr\left(v_{ij} = 1\right) = 0{,}2 \,\forall j = 3, 6 \wedge i = 1, \ldots, 25 \tag{A.1}$$

  Für jedes Objekt i wurden zwei auf dem Intervall $[0; 1]$ stetige, gleichverteilte Zufallszahlen generiert. War die gezogene Zufallszahl kleiner gleich als der in dem Modell A.1 angegebene Wert, so wurde die zugehörige Merkmalsausprägung a_{i3} bzw. a_{i6} als fehlend betrachtet. Die auf diese Art ermittelten Werte sind in der Tabelle A.1 grau hinterlegt und unterstrichen.

- **Fall 2:**
  Die Daten fehlen MAR. In Abhängigkeit der in dem Merkmal SEX vorhandenen Ausprägung unterscheidet sich die Ausfallwahrscheinlichkeit im Merkmal AGE. So ist für jede Person i ($i = 1, \ldots$,25) die Wahrscheinlichkeit, dass die Ausprägung a_{i4} fehlt, gleich

$$Pr\left(v_{i4} = 1 | a_{i3}\right) = 0{,}2 \cdot \left(g\left(a_{i3}\right) + 1\right)$$
$$\text{mit } g\left(a_{i3}\right) = \begin{cases} 1 & \text{falls } a_{i3} = M \\ 0 & \text{falls } a_{i3} = F \end{cases} \tag{A.2}$$

  ist. Die auf diese Art ermittelten Werte sind in der Tabelle A.1 fett formatiert und unterstrichen.

- **Fall 3:**
  Die Daten fehlen MAR. In Abhängigkeit der Merkmalsausprägung von EDUCD steigt die Ausfallwahrscheinlichkeit bei INCTOT. Für die Abhängigkeit wird ein Logit-Modell der folgenden Form unterstellt:

$$Pr\left(v_{i9}=1|a_{i5}\right)=\frac{\exp(-6+0{,}25\cdot a_{i5})}{1+\exp(-6+0{,}25\cdot a_{i5})}\;\forall i=1,\ldots,25 \tag{A.3}$$

  Für jedes Objekt i wurde eine auf dem Intervall $[0;1]$ stetige, gleichverteilte Zufallszahl generiert. War die gezogene Zufallszahl kleiner als die sich für das Objekt i aus dem Modell A.1 ergebende Ausfallwahrscheinlichkeit, so wurde die Merkmalsausprägung a_{i9} als fehlend betrachtet. Die auf diese Art ermittelten Werte sind in der Tabelle A.1 bei dem betroffenen Merkmal grau hinterlegt.

- **Fall 4:**
  Die Daten fehlen NMAR. In Abhängigkeit der in dem Merkmal FTOTINC vorhandenen Ausprägung unterscheidet sich die Ausfallwahrscheinlichkeit im Merkmal FTOTINC. Maßgeblich für den Ausfall ist folgendes lineares Wahrscheinlichkeitsmodell:

$$Pr\left(v_{i10}=1|a_{i10}\right)=\begin{cases}0 & \text{falls } a_{i10}<25.000\\ \frac{0{,}6}{225.000}\cdot a_{i10} & \text{falls } a_{i3}\in[25.000{,}250.000]\\ 0{,}6 & \text{falls } a_{i10}>250.000\end{cases} \tag{A.4}$$

  Die auf diese Art ermittelten Werte sind in der Tabelle A.1 bei dem betroffenen Merkmal grau hinterlegt.

Objekt	PERWT	REGION	SEX	AGE	EDUCD	LANGUAGE	TRANTIME	UHRSWORK	INCTOT	FTOTINC	INCSS
1	66	S	M	19	11	Eng	5	40	13.000	57.000	0
2	68	S	F	56	5	Sp	30	40	11.000	46.000	0
3	215	N	F	**58**	9	Eng	0	0	1.820	1.820	0
4	101	C	M	61	12	Eng	0	0	40.000	**40.000**	0
5	81	C	M	61	9	Eng	0	0	7.500	**60.500**	7.500
6	83	C	M	**66**	16	Eng	15	24	249.000	258.100	24.000
7	80	W	M	16	9	Eng	0	0	0	**119.000**	0
8	294	W	F	**86**	15	Fr	0	0	5.000	**159.800**	0
9	177	S	M	50	13	Sp	15	40	65.000	116.000	0
10	90	W	F	**93**	12	Ger	0	0	13.600	28.600	10.600
11	59	S	F	**29**	15	Eng	60	40	30.000	87.000	0
12	76	S	F	56	12	Eng	10	10	7.700	28.400	0
13	247	N	M	**48**	12	Eng	15	40	18.000	**60.130**	0
14	132	S	M	28	13	Eng	15	40	22.000	22.000	0
15	136	S	F	66	15	Eng	15	10	11.600	**34.600**	11.200
16	357	W	M	**27**	12	Sp	45	40	11.500	176.500	0
17	206	C	F	77	13	Eng	0	0	9.100	9.100	3.100
18	101	S	M	50	17	Eng	30	40	58.000	58.000	0
19	383	W	M	43	12	Eng	0	0	1.300	8.500	0
20	97	S	M	65	15	Eng	0	50	74.000	**122.000**	0
21	247	S	M	23	13	Eng	0	0	0	0	0
22	236	S	M	30	15	Sp	0	40	42.004	184.104	0
23	91	S	F	54	13	Eng	60	40	35.000	**270.000**	0
24	88	S	F	70	12	Fr	0	0	34.400	50.600	15.000
25	260	N	M	33	13	Eng	30	40	21.000	21.000	0

Tabelle A.1: Stichprobe der 2009 ACS

Anhang B

Weitere Ausführungen zu den Distanzmaßen

B.1 Spezialfälle der Minkowski-Distanz

Eine Veränderung in der Bewertung von Unterschieden in den Merkmalsausprägungen, die sich durch eine Variation des Parameters p ergeben lässt sich für den Zwei-Merkmal-Fall grafisch veranschaulichen. Die in Abbildung B.1 dargestellten Linien entsprechen jenen Punkten, die für eine Auswahl an p, jeweils eine Distanz von Eins für den Zwei-Merkmal-Fall aufweisen[1]. Im Folgenden werden drei Spezialfälle[2] erörtert, für die sich nicht nur in der Missing-Data-Literatur spezielle Namen etabliert haben.

B.1.1 Die Manhattan-Distanz

Für $p = 1$ ergibt sich folgender Spezialfall mittels (3.1):

$$c_{ij} = \sum_{k=1}^{m-q} \alpha_k \cdot |a_{ik}^{cv} - a_{jk}^{cv}| \quad \forall\, i \in D,\, j \in R. \tag{B.1}$$

Der Name Manhattan-Distanz oder auch City-Block-Metrik kommt daher, dass die Block-Struktur des Bebauungsplans von Manhattan Autofahrer dazu

[1] Auf die Darstellung der verbleibenden Quadranten wurde aufgrund der Achsensymmetrie der Äquidistanzlinien verzichtet.

[2] Diese Spezialfälle ergeben sich für $p = 1$, $p = 2$ und $p \to \infty$.

zwingt, zwischen zwei Punkten in der Stadt immer den vollständigen Längs- und Breitenweg zurückzulegen.

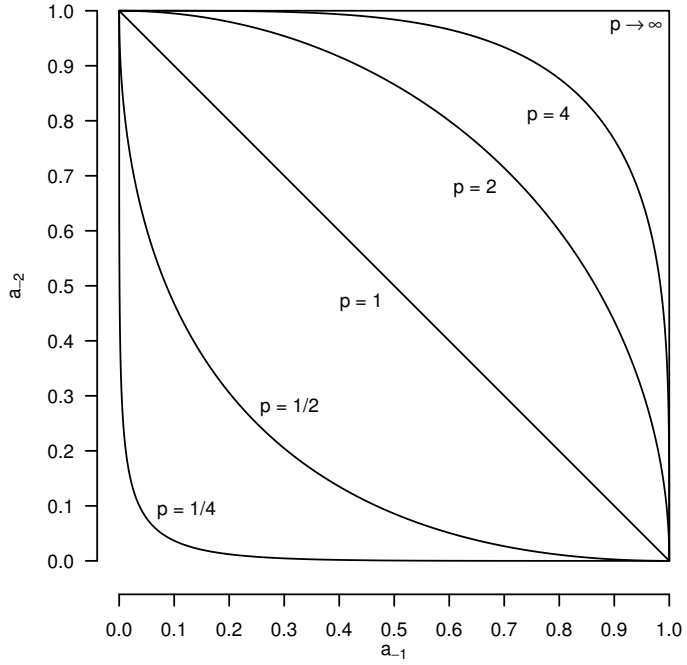


Abbildung B.1: Punkte mit einer Distanz von Eins zum Ursprung unter Variation des Parameters p, zwei Merkmale

Im Bereich der Hot-Deck-Imputation findet die Manhattan-Distanz beispielsweise im Canadian Census of Construction (vgl. Ford, 1983, S. 199 f., 204) Anwendung. Ferner wird sie von Yenduri und Iyengar (2007, S. 136) in einer Simulationsstudie mit anderen Metriken verglichen und von de Waal et al. (2011, S. 408) in einer Beispielrechnung verwendet.

Beispiel B.1: *Distanzberechnung mittels der Manhattan-Distanz*
Für die Datenmatrix des Anhangs A soll im Folgenden der Fall 3 betrachtet werden. Dabei werden die Manhattan-Distanzen zwischen den ersten beiden Spender-Empfänger-Kombinationen unter der Verwendung der Merkmale AGE, TRANTIME und UHRSWORK berechnet. Als merkmalsspezifische Gewichte werden die Inversen der Merkmalsspannweiten verwendet, die unter der Verwendung aller vorhandenen Merkmalsausprägungen bestimmt wurden. Hierdurch kann jeder Summand der Distanz maximal den Wert eins annehmen.

Dies resultiert in den folgenden beiden Distanzen:

$$c_{21} = \frac{|56-19|}{93-16} + \frac{|30-5|}{60-0} + \frac{|40-40|}{50-0} = 0{,}48 + 0{,}41 + 0 = 0{,}89$$
$$c_{31} = \frac{|58-19|}{93-16} + \frac{|0-5|}{60-0} + \frac{|0-40|}{50-0} = 0{,}50 + 0{,}08 + 0{,}80 = 1{,}38.$$

B.1.2 Die Euklidischen-Distanz

Mit der Wahl von $p = 2$ ergibt sich aus (3.1) folgender Spezialfall:

$$c_{ij} = \sqrt{\sum_{k=1}^{m-q} \alpha_k \cdot |a_{ik}^{cv} - a_{jk}^{cv}|^2} \;\; \forall\, i \in D,\, j \in R. \tag{B.2}$$

Die Möglichkeit, diese Euklidische-Distanz zu verwenden, um einem Empfänger ähnliche Spender zu bestimmen, wurde bereits von (Ford, 1983, S. 187) vorgeschlagen. Im 2002 Agriculture Census der USA wird sie, unter der Verwendung von Merkmalen wie Gesamtnutzfläche, Längen- und Breitengrad, zur Imputation fehlender Werte bei befragten Landwirtschaftsbetrieben verwendet (Hogye, 2004, S. 3672). Auch in Simulationsstudien, die Auswirkungen verschiedener Imputationsverfahren auf die Güte der Imputationsergebnisse untersuchen, wird die Euklidische-Distanz in Hot-Deck-Verfahren angewandt (vgl. Roth et al., 1999, S. 212 und Yenduri und Iyengar, 2007, S. 136). Huisman (2000, S. 335) verwendet hingegen eine quadrierte Euklidische-Distanz.

Beispiel B.2: *Distanzberechnung mittels der Euklidische-Distanz*
Die exemplarische Berechnung der Euklidischen-Distanz soll im Folgenden anhand des Falls 3 der in Anhang A gegebenen Datenmatrix erfolgen. Die ersten beiden Spender-Empfänger-Distanzen werden auch in diesem Beispiel mittels der quantitativen Merkmale AGE, TRANTIME und UHRSWORK berechnet, jedoch entsprechen dieses Mal die α_k dem Kehrwert der Merkmalsstichprobenvarianzen. Diese Gewichtung entspricht einer vorherigen Skalierung der Merkmale, so dass diese eine Standardabweichung von Eins aufweisen. Diese

Vorgehensweise resultiert in den folgenden Distanzen:

$$\begin{aligned} c_{21} =& \sqrt{\frac{|56-19|^2}{434{,}91} + \frac{|30-5|^2}{350{,}58} + \frac{|40-40|^2}{394{,}57}} \\ =& \sqrt{3{,}14 + 1{,}78 + 0} = 2{,}22 \\ c_{31} =& \sqrt{\frac{|58-19|^2}{434{,}91} + \frac{|0-5|^2}{350{,}58} + \frac{|0-40|^2}{394{,}57}} \\ =& \sqrt{3{,}49 + 0{,}07 + 4{,}05} = 2{,}76. \end{aligned}$$

B.1.3 Die Tschebyscheff-Distanz

In der gewichteten L_p-Distanz steigt der Anteil der größten Summanden von c_{ij} an der Gesamtsumme mit steigendem p. Geht p gegen unendlich[3], so wird c_{ij} lediglich durch den größten Summanden definiert und kann in folgender Form geschrieben werden:

$$\begin{aligned} c_{ij} &= \lim_{p\to\infty} \left(\sum_{k=1}^{m-q} \alpha_k \cdot |a_{ik}^{cv} - a_{jk}^{cv}|^p \right)^{\frac{1}{p}} \\ &= \max_{k\in\{1,\ldots,m-q\}} \left(\alpha_k \cdot |a_{ik}^{cv} - a_{jk}^{cv}| \right) \quad \forall\, i \in D,\, j \in R. \end{aligned} \tag{B.3}$$

Dieser Spezialfall ist gemeinhin unter dem Namen Tschebyscheff-Distanz oder auch Maximum-Norm bekannt und wurde im Rahmen von Hot-Deck-Verfahren bereits von Sande (1983, S. 344) erwähnt. Auch in den neueren Werken von Little und Rubin (2002, S. 69) sowie Andridge und Little (2010, S. 44) findet diese Metrik bei der Beschreibung von Hot-Deck-Verfahren Erwähnung.

Die Maximum-Norm war Grundlage der Spender-Identifizierung in dem für Statistics Canada entwickelten Generalized Edit and Imputation System (GEIS) (Rancourt, 1999, S. 132). Auch der Nachfolger von GEIS, Banff System for Edit and Imputation, verwendet diese Metrik (Kozak, 2005, S. 6). Konkrete Anwendung findet die Tschebyscheff-Distanz in der Unified Enterprise Survey (vgl. Whitridge und Kovar, 1990, S. 105 ff.) sowie der Survey of Household Spending (SHS) (Rancourt, 1999, S. 132) und Financial Farm Survey (FFS) (Rancourt, 1999, S. 133).

[3] Des Weiteren gilt $\lim_{p\to\infty} 1/p = 0$.

Beispiel B.3: *Distanzberechnung mittels der Tschebyscheff-Distanz*
Für die ersten beiden Spender für den ersten Empfänger des Falls 3 der im Anhang A gegebenen Datenmatrix werden nun die Tschebyscheff-Distanzen berechnet. Zur Berechnung dieser Maximum-Norm werden die vollständigen Kovariaten AGE, TRANTIME und UHRSWORK herangezogen, wobei für die merkmalsspezifischen Gewichte wiederum die Kehrwerte der Merkmalsspannweiten gewählt werden. Es entstehen folgende Spender-Empfänger Distanzen:

$$\begin{aligned} c_{21} &= \max\left\{\frac{|56-19|}{93-16}; \frac{|30-5|}{60-0}; \frac{|40-40|}{50-0}\right\} \\ &= \max\{0{,}48; 0{,}41; 0\} = 0{,}48 \\ c_{31} &= \max\left\{\frac{|58-19|}{93-16}; \frac{|0-5|}{60-0}; \frac{|0-40|}{50-0}\right\} \\ &= \max\{0{,}50; 0{,}08; 0{,}80\} = 0{,}80. \end{aligned}$$

B.2 Beispiele zur Mahalanobis-Distanz

Beispiel B.4: *Distanzberechnung mittels der Mahalanobis-Distanz*
Im Folgenden werden zwei Mahalanobis-Distanzen exemplarisch für den im Anhang A zu findenden Fall 3 bestimmt. Die zwei Distanzen werden unter der Verwendung der Objekte mit den Indizes eins, zwei und drei, also dem ersten Empfänger und den ersten beiden Spendern, berechnet. Wiederum werden zur Berechnung die drei vollständig vorhandenen quantitativen Merkmale AGE, TRANTIME und UHRSWORK verwendet. Zur Berechnung sind die folgende Kovarianzmatrix und die darauf basierenden Inversen erforderlich:

$$\Sigma^{cv} = \begin{pmatrix} 434{,}91 & -106{,}95 & -166{,}93 \\ -106{,}95 & 350{,}58 & 226{,}70 \\ -166{,}93 & 226{,}70 & 394{,}57 \end{pmatrix}$$

$$\Sigma^{cv-1} \approx \begin{pmatrix} 2{,}74 & 0{,}13 & 1{,}08 \\ 0{,}13 & 4{,}54 & -2{,}55 \\ 1{,}08 & -2{,}55 & 4{,}45 \end{pmatrix} \cdot 10^{-3}.$$

Damit ergibt sich folgende quadrierte Mahalanobis-Distanz zwischen den Objekten zwei und eins:

$$c_{21}^2 = \left(\begin{pmatrix} 56 \\ 30 \\ 40 \end{pmatrix} - \begin{pmatrix} 19 \\ 5 \\ 40 \end{pmatrix}\right)^T \Sigma^{cv-1} \left(\begin{pmatrix} 56 \\ 30 \\ 40 \end{pmatrix} - \begin{pmatrix} 19 \\ 5 \\ 40 \end{pmatrix}\right)$$
$$\approx (0{,}105\ 0{,}118\ -0{,}023) \begin{pmatrix} 37 \\ 25 \\ 0 \end{pmatrix} = 6{,}859$$

sowie die quadrierte Distanz zwischen den Objekten drei und eins:

$$c_{31}^2 = \left(\begin{pmatrix} 58 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 19 \\ 5 \\ 40 \end{pmatrix}\right)^T \Sigma^{cv-1} \left(\begin{pmatrix} 58 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 19 \\ 5 \\ 40 \end{pmatrix}\right)$$
$$\approx (0{,}063\ 0{,}0847\ -0{,}123) \begin{pmatrix} 39 \\ -5 \\ -40 \end{pmatrix} = 6{,}974.$$

Zuletzt ist die Wurzel aus beiden quadrierten Distanzen zu ziehen, was zu den Distanzen $c_{21} = 2{,}619$ und $c_{31} = 2{,}640$ führt.

Beispiel B.5: *Berechnung einer zielgerichteten Mahalanobis-Distanz*
Ausgehend von der in Beispiel B.4 berechneten Kovarianzmatrix ergeben sich durch die Lösung des entsprechenden Eigenwertproblems die folgenden Eigenwerte von Σ^{cv}:

$$\lambda_1 = 730{,}47; \lambda_2 = 309{,}91; \lambda_3 = 139{,}68$$

sowie die zugehörige Matrix der Eigenvektoren

$$\Phi \approx \begin{pmatrix} 0{,}553 & -0{,}818 & 0{,}153 \\ -0{,}536 & -0{,}491 & -0{,}686 \\ -0{,}637 & -0{,}297 & 0{,}710 \end{pmatrix}.$$

Mittels dieser lässt sich nun b_{1-} wie folgt berechnen:

$$b_{1-} = \Phi^T a_{1-} = \begin{pmatrix} 0{,}553 & -0{,}536 & -0{,}637 \\ -0{,}818 & -0{,}491 & -0{,}297 \\ 0{,}153 & -0{,}686 & 0{,}710 \end{pmatrix} \begin{pmatrix} 19 \\ 5 \\ 40 \end{pmatrix} = \begin{pmatrix} -17{,}640 \\ -29{,}925 \\ 27{,}915 \end{pmatrix}.$$

Analog lässt sich mit $b_{2-} = (-10{,}551\ -72{,}490\ 16{,}426)^T$ und
$b_{3-} = (32{,}122\ -47{,}466\ 8{,}890)^T$ die gesamte Matrix B berechnen.

Werden nun die Euklidischen-Distanzen c_{21} bzw. c_{31} mit $\alpha_k = 1/\lambda_k$ berechnet, entstehen wieder die aus Beispiel B.4 bekannten Mahalanobis-Distanzen:

$$\begin{aligned} c_{21} =& \sqrt{\frac{|-10{,}551+17{,}640|^2}{730{,}47} + \frac{|-72{,}490+29{,}925|^2}{309{,}91} + \frac{|16{,}426-27{,}915|^2}{139{,}68}} \\ =& \sqrt{0{,}068+5{,}845+0{,}944} = 2{,}619 \\ c_{31} =& \sqrt{\frac{|32{,}12+17{,}640|^2}{730{,}47} + \frac{|-47{,}46+29{,}925|^2}{309{,}91} + \frac{|8{,}89-27{,}915|^2}{139{,}68}} \\ =& \sqrt{3{,}390+0{,}992+2{,}591} = 2{,}640. \end{aligned}$$

Werden hingegen die Korrelationen mit einer Complete-Case-Analyse zwischen den drei Hauptkomponenten und der in Fall 3 zu imputierenden Variable INCTOT berücksichtigt, ergeben sich folgende veränderte Gewichte:

$$\begin{aligned} \alpha_1 =& \frac{|-0{,}028|}{730{,}47} \\ \alpha_2 =& \frac{|-0{,}305|}{309{,}91} \\ \alpha_3 =& \frac{|0{,}283|}{139{,}68}. \end{aligned}$$

Die Korrelationen verdeutlichen, dass die erste Hauptkomponente einen geringen Zusammenhang mit den vorhandenen Werten des Merkmals INCTOT aufweist. Differenzen in dieser Hauptkomponente werden daher weniger berücksichtigt als Differenzen in den anderen Merkmalen. Mittels dieser Gewichte können nun die zielgerichteten Mahalanobis-Distanzen wie folgt berechnet werden:

$$\begin{aligned} c_{21} =& \sqrt{0{,}028\frac{|7{,}089|^2}{730{,}47} + 0{,}305\frac{|42{,}564|^2}{309{,}91} + 0{,}283\frac{|11{,}488|^2}{139{,}68}} \\ =& \sqrt{2{,}64\cdot 10^{-6} + 0{,}00576 + 0{,}00191} = 0{,}0876 \\ c_{31} =& \sqrt{0{,}028\frac{|49{,}762|^2}{730{,}47} + 0{,}305\frac{|17{,}541|^2}{309{,}91} + 0{,}283\frac{|19{,}025|^2}{139{,}68}} \\ =& \sqrt{0{,}000130+0{,}000979+0{,}0052} = 0{,}0798. \end{aligned}$$

Anhand dieses Beispiels wird bereits deutlich, dass die Objekte zwei und eins zwar an sich ähnlicher sind als die Objekte drei und eins, jedoch kann sich diese Beziehung umkehren, wenn ein Bezug auf jenes Merkmal genommen wird, welches imputiert werden soll.

B.3 Weitere Distanzfunktionen

In der Literatur finden weitere Distanzfunktionen im Zusammenhang mit Hot-Deck-Verfahren Erwähnung. Im Wesentlichen sind diese entweder der Simulationsstudie von Yenduri und Iyengar (2007, S. 136 f.) oder den Arbeiten von Rubin et al. (Rubin, 1979; Rosenbaum und Rubin, 1983; Rosenbaum und Rubin, 1985; Rubin und Thomas, 1992) entnommen und finden in anderen Arbeiten selten Erwähnung beziehungsweise Anwendung. Hierbei handelt es sich um folgende Distanzen:

B.3.1 Korrelation-Distanz

Yenduri und Iyengar (2007, S. 136) verwenden die Korrelation zwischen zwei Objekten als Ähnlichkeitsmaß[4]. Eine Größe, welche aus diesem Ähnlichkeitsmaß abgeleitet werden kann, auf den Bereich zwischen Null und Eins normiert ist und die Eigenschaften einer Distanz aufweist, kann wie folgt berechnet werden:

$$c_{ij} = \frac{1}{2} - \frac{\sum_{k=1}^{m-q} (a_{ik}^{cv} - \bar{a}_{i\bullet}^{cv})(a_{jk}^{cv} - \bar{a}_{j\bullet}^{cv})}{2 \cdot \sqrt{\sum_{k=1}^{m-q} (a_{ik}^{cv} - \bar{a}_{i\bullet}^{cv})^2 \cdot \sum_{k=1}^{m-q} (a_{jk}^{cv} - \bar{a}_{j\bullet}^{cv})^2}} \quad \forall\, i \in D,\, j \in R, \tag{B.4}$$

wobei $\bar{a}_{i\bullet}^{cv}$ beziehungsweise $\bar{a}_{j\bullet}^{cv}$ die Mittelwerte des i-ten Spenders beziehungsweise j-ten Empfängers, berechnet über die Ausprägungen der vollständig vorhandenen Hilfsvariablen, sind.

Beispiel B.6: *Berechnung einer Korrelation-Distanz*
Zur exemplarischen Berechnung zweier Korrelation-Distanzen werden die ersten Spender und der erste Empfänger der Datenmatrix des Anhangs A betrachtet. Als Hilfsvariablen dienen die vollständigen Merkmale AGE, TRANTIME und UHRSWORK. Zunächst sind zur Berechnung dieser Distanzen die Mittelwerte über alle Merkmalsausprägungen der Objekte eins bis drei zu berechnen. Diese sind $\bar{a}_{1\bullet} = 21{,}33$ für den Empfänger und $\bar{a}_{2\bullet} = 42$ beziehungsweise $\bar{a}_{3\bullet} = 19{,}33$ für die beiden Spender. Danach erfolgen die Berechnung der Varianzen der Objekte sowie der Kovarianzen der Spender zu dem Empfänger.

[4] Die Autoren bezeichnen dieses Ähnlichkeitsmaß irrtümlich als Distanz.

Anschließend werden die Korrelation-Distanzen wie folgt berechnet:

$$c_{21} = \frac{1}{2} - \frac{63}{2 \cdot \sqrt{172 \cdot 310{,}33}} = 0{,}363$$
$$c_{31} = \frac{1}{2} - \frac{-67{,}66}{2 \cdot \sqrt{1.121{,}33 \cdot 310{,}33}} = 0{,}557.$$

B.3.2 Kosinus-Distanz

Ein weiteres Ähnlichkeitsmaß[5], das in der Studie von Yenduri und Iyengar (2007, S. 137) verwendet wird, basiert auf der Berechnung des Kosinus des Winkels zwischen den Spendern und Empfängern. Auch diese Ähnlichkeit kann entsprechend transformiert werden, so dass die resultierende Größe die Eigenschaften einer Distanz aufweist. Nach einer entsprechenden Transformation lässt sich eine Kosinus-Distanz wie folgt berechnen:

$$c_{ij} = \frac{1}{2} - \frac{\sum_{k=1}^{m-q} a_{ik}^{cv} \cdot a_{jk}^{cv}}{2 \cdot \sqrt{\sum_{k=1}^{m-q} (a_{ik}^{cv})^2 \cdot \sum_{k=1}^{m-q} (a_{jk}^{cv})^2}} \; \forall\, i \in D,\, j \in R. \tag{B.5}$$

Beispiel B.7: *Berechnung einer Kosinus-Distanz*
In der folgenden Beispielrechnung wird für den Fall 3 der Datenmatrix des Anhangs A die Kosinus-Distanz zwischen dem ersten Empfänger und den beiden ersten Spendern ermittelt. Zur Berechnung der Distanzen werden die Ausprägungen der entsprechenden Objekte bei den Merkmalen AGE, TRANTIME und UHRSWORK herangezogen. Die entsprechenden Spender-Empfänger-Distanzen berechnen sich wie folgt:

$$c_{21} = \frac{1}{2} - \frac{56 \cdot 19 + 30 \cdot 5 + 40 \cdot 40}{2 \cdot \sqrt{(56^2 + 30^2 + 40^2) \cdot (19^2 + 5^2 + 40^2)}} = 0{,}079$$
$$c_{31} = \frac{1}{2} - \frac{58 \cdot 19 + 0 \cdot 5 + 0 \cdot 40}{2 \cdot \sqrt{(58^2 + 0^2 + 0^2) \cdot (19^2 + 5^2 + 40^2)}} = 0{,}286$$

B.3.3 Squared-Chord-Distanz

Die letzte von Yenduri und Iyengar (2007, S. 137) verwendete Distanz wird Squared-Chord-Distanz genannt und berechnet sich folgendermaßen:

$$c_{ij} = \sum_{k=1}^{m-q} \left(\sqrt{a_{ik}^{cv}} - \sqrt{a_{jk}^{cv}} \right)^2 \; \forall\, i \in D\, j \in R. \tag{B.6}$$

[5] Wiederum irrtümlicher Weise als Distanz bezeichnet.

Es ist offensichtlich, dass diese Formel lediglich für nicht negative Werte Anwendung finden kann. Dieses Anforderung kann durch die Addition des entsprechenden Merkmalsminimums zu jeglichen Merkmalsausprägungen behoben werden (vgl. Yenduri und Iyengar, 2007, S. 137).

Beispiel B.8: *Berechnung einer Squared-Chord-Distanz*
Um die Berechnung der Squared-Chord-Distanz an dieser Stelle exemplarisch darzulegen, werden die Merkmale AGE, TRANTIME und UHRSWORK der Datenmatrix im Fall 3 des Anhangs A verwendet. Ermittelt werden wieder die Distanzen zwischen dem ersten Empfänger und den ersten beiden Spendern. Da alle Werte nicht negativ sind, muss keine Korrektur vorgenommen werden. Es ergeben sich folgende Berechnungen:

$$\begin{aligned} c_{21} &= \left(\sqrt{56}-\sqrt{19}\right)^2 + \left(\sqrt{30}-\sqrt{5}\right)^2 + \left(\sqrt{40}-\sqrt{40}\right)^2 \quad = 20{,}267 \\ c_{31} &= \left(\sqrt{58}-\sqrt{19}\right)^2 + \left(\sqrt{0}-\sqrt{5}\right)^2 + \left(\sqrt{0}-\sqrt{40}\right)^2 \quad = 55{,}607. \end{aligned}$$

B.3.4 Discriminant-Matching

Discriminant-Matching (Rubin, 1976b; Rubin, 1980[6] beziehungsweise Rubin, 1979, S. 319; Rubin und Thomas, 1992, S. 798 ff.) wurde zuerst von Rubin (1987, S. 158) zur Nutzung im Kontext von Hot-Deck-Verfahren vorgeschlagen. Gemäß Rubin (1987, S. 158) werden in Beobachtungsstudien Objekte aus den Kontroll- und Experimentalgruppen einander zugewiesen, analog zu der Spender-Empfänger-Zuordnung bei Hot-Deck-Verfahren. Daher könnte auch bei den Hot-Deck-Verfahren das Discriminant-Matching eingesetzt werden, welches grundsätzlich auf einer Gruppenzugehörigkeitsprädiktion basiert. Beim Discriminant-Matching wird der Wert jedes Objekts auf der geschätzten Diskriminante zwischen den Spendern und Empfängern auf den Kovariaten berechnet. Demnach wird $A^{cv}\mathfrak{D}$ zur Zuordnung verwendet, wobei $\mathfrak{D} = \Sigma^{cv-1}\left(\bar{a}^{cv}_{D-} - \bar{a}^{cv}_{R-}\right)$ und $\bar{a}^{cv}_{D-}$ beziehungsweise $\bar{a}^{cv}_{R-}$ den Spaltenvektoren der Merkmalsmittelwerte bezüglich aller Objekte in D beziehungsweise R entspricht. Dies kann auch äquivalent als Distanz dargestellt werden (vgl. Rubin, 1979, S. 319):

$$c_{ij} = \left(a^{cv}_{i-} - a^{cv}_{j-}\right)^T \mathfrak{D}\mathfrak{D}^T \left(a^{cv}_{i-} - a^{cv}_{j-}\right) \;\forall\, i \in D,\, j \in R. \tag{B.7}$$

[6] Zitiert nach Rubin (1987, S. 158).

Beispiel B.9: *Distanzberechnung beim Discriminant-Matching*
Für die Datenmatrix des Anhangs A wird im Folgenden der Fall 3 mit den vollständigen Merkmalen AGE, TRANTIME und UHRSWORK betrachtet. Die Berechnung von $\mathfrak{D}$ erfordert die Kovarianzmatrix und den Mittelwertsvektor der vollständigen Merkmale für jeweils die Spender und die Empfänger. Σ^{cv-1} entspricht jener in Beispiel B.4 verwendeten Kovarianzmatrix, so dass sich $\mathfrak{D}$ wie folgt berechnet:

$$\begin{aligned}\mathfrak{D} &= \Sigma^{cv-1}\left((51{,}38\ 13{,}33\ 22{,}44)^T - (48{,}57\ 15{,}00\ 18{,}57)^T\right)\\ &= (0{,}011\ -0{,}017\ 0{,}024)^T\,.\end{aligned}$$

Mit Hilfe von $\mathfrak{D}$ und der Formel (B.7) können nun Spender-Empfänger-Distanzen ermittelt werden:

$$\begin{aligned}c_{21} &= \left(\begin{pmatrix}56\\30\\40\end{pmatrix} - \begin{pmatrix}19\\5\\40\end{pmatrix}\right)^T \mathfrak{D}\mathfrak{D}^T \left(\begin{pmatrix}56\\30\\40\end{pmatrix} - \begin{pmatrix}19\\5\\40\end{pmatrix}\right)\\ &= \begin{pmatrix}37\\25\\0\end{pmatrix}^T \begin{pmatrix}1{,}37 & -2{,}00 & 2{,}87\\ -2{,}00 & 2{,}91 & -4{,}19\\ 2{,}87 & -4{,}19 & 6{,}04\end{pmatrix} \begin{pmatrix}37\\25\\0\end{pmatrix} \cdot 10^{-4} \quad = 4{,}24 \cdot 10^{-5}\end{aligned}$$

$$\begin{aligned}c_{31} &= \left(\begin{pmatrix}58\\0\\0\end{pmatrix} - \begin{pmatrix}19\\5\\40\end{pmatrix}\right)^T \mathfrak{D}\mathfrak{D}^T \left(\begin{pmatrix}58\\0\\0\end{pmatrix} - \begin{pmatrix}19\\5\\40\end{pmatrix}\right)\\ &= \begin{pmatrix}39\\-5\\-40\end{pmatrix}^T \begin{pmatrix}1{,}37 & -2{,}00 & 2{,}87\\ -2{,}00 & 2{,}91 & -4{,}19\\ 2{,}87 & -4{,}19 & 6{,}04\end{pmatrix} \begin{pmatrix}39\\-5\\-40\end{pmatrix} \cdot 10^{-4} \quad = 0{,}194.\end{aligned}$$

B.4 Beispiel zum Vergleich der Kodierungen

Beispiel B.10: *Vergleich der Kodierungen für die Berechnung von Distanzen*
Für die Datenmatrix des Anhangs A wird im Folgenden betrachtet, wie sich unterschiedliche Kodierungsverfahren auf eine Distanz auswirken, die rein auf das Merkmal REGION berechnet werden. Zur Distanzberechnung wird eine

ungewichtete Manhattan-Distanz verwendet, welche im Falle der Dummy-Kodierung dieselben Ergebnisse liefert wie eine Zählung der nicht-übereinstimmenden Merkmale. Zur Kodierung sind drei beziehungsweise vier neue Merkmale erforderlich, wenn keine Referenzkategorie festgelegt wird. Eine Dummy-Codierung der Merkmalsausprägungen für die ersten vier Objekte[7], ohne die Verwendung einer Referenzkategorie, liefern folgende abgewandelte Datenmatrix und die zugehörige vollständige Distanzmatrix:

$$A' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad C' = \begin{pmatrix} 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 \\ 2 & 2 & 0 & 2 \\ 2 & 2 & 2 & 0 \end{pmatrix}.$$

Wird hingegen bei der Dummy-Kodierung die Kategorie S als Referenzkategorie festgelegt, entstehen die folgenden abweichenden Ergebnisse:

$$A'' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad C'' = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 1 & 1 & 2 & 0 \end{pmatrix}.$$

Bei einer Effekt-Kodierung, bei der auch die Kategorie S als Referenzkategorie festgelegt wird, entstehen folgende Daten- und Distanzmatrizen:

$$A''' = \begin{pmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad C''' = \begin{pmatrix} 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 4 & 4 & 0 & 2 \\ 4 & 4 & 2 & 0 \end{pmatrix}.$$

Anhand der Datenmatrizen ist deutlich zu erkennen, dass die Verwendung einer Referenzkategorie zu einer kompakteren Darstellungsform führt. Kritisch anzumerken ist jedoch, sofern das nominale Merkmal nicht dichotom ist, dass die Verwendung einer Referenzkategorie bei der Dummy-Kodierung und die Effekt-Kodierung zu Verzerrungen in der Distanzmatrix führen. Bei einer Dummy-Codierung mit Referenzkategorie werden die Distanzen zu jenen Objekten, bei denen die Referenzkategorie als Merkmalsausprägung auftritt, immer kleiner sein als die Distanzen von Objekten, bei denen bei beiden nicht die Referenzkategorie als Merkmalsausprägung auftritt (vgl. c''_{24} und c''_{34}). Bei einer

[7] Die Ausprägungen der ersten vier Objekte des Merkmals REGION lauten $(S\ S\ N\ C)^T$.

Effekt-Codierung werden die entsprechenden Distanzen immer größer sein. Diese Verzerrungen sind als unerwünscht zu betrachten, da die Referenzkategorie willkürlich festgelegt werden kann und somit grundsätzlich jede Kategorie hierzu dienen kann. Somit liefert die Verwendung einer Dummy-Codierung ohne Referenzkategorie für die Zwecke einer Distanzberechnung bessere Ergebnisse, da jede Nichtübereinstimmung gleich gewertet wird. Bei dichotomen Merkmalen treten derartige Unterschiede bei den Distanzen nicht auf, so dass hier keine der drei Kodierungsverfahren grundsätzlich besser geeignet ist als eine andere.

Symbolverzeichnis

Symbol	Bedeutung
$\circ$	Andeutungssymbol für fehlende Werte
$\lceil \bullet \rceil$	Aufrundungsfunktion
$\mid \bullet \mid$	Betragsfunktion bzw. die Mächtigkeit einer Menge
α_k	merkmalsspezifisches Gewicht
α_{kl}	Verhältnis (ratio) zwischen Merkmal k und Merkmal l
$A = (a_{ik})_{n,m} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix}$	Datenmatrix mit n Objekten und m Merkmalen
$a_{i-} = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{im} \end{pmatrix}$ bzw. $a_{-k} = \begin{pmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{pmatrix}$	Objektvektor i bzw. Merkmalsvektor k der Datenmatrix A
a_{ik}	Ausprägung des Merkmals k bei Objekt i
$\hat{a}_{ik}$	Schätzwert für die Ausprägung des Merkmals k bei Objekt i
$\bar{a}_{i\bullet}$ bzw. $\bar{a}_{\bullet k}$	Mittelwert der Merkmalsausprägungen von Objekt i bzw. Merkmal k
A^{obs}	vorhandener Teil der Ausprägungen aus A

Symbol	Bedeutung
a_{i-}^{obs} bzw. a_{-k}^{obs}	vorhandener Teil der Ausprägungen aus a_{i-} bzw. a_{-k}
A^{mis}	fehlender Teil der Ausprägungen aus A
a_{i-}^{mis} bzw. a_{-k}^{mis}	fehlender Teil der Ausprägungen aus a_{i-} bzw. a_{-k}
$A^{cc} = (a_{ik}^{cc})_{n-r,m}$	Teilmatrix von A jener Objekte, bei denen keine Merkmalsausprägung fehlt
a_{i-}^{cc} bzw. a_{-k}^{cc}	Objektvektor i bzw. Merkmalsvektor k der Matrix A^{cc}
$A^{mc} = (a_{ik}^{mc})_{r,m}$	Teilmatrix von A jener Objekte, bei denen mindestens eine Merkmalsausprägung fehlt
a_{i-}^{mc} bzw. a_{-k}^{mc}	Objektvektor i bzw. Merkmalsvektor k der Datenmatrix A^{mc}
$A^{cv} = (a_{ik}^{cv})_{n,m-q}$	Teilmatrix von A jener Merkmale, bei denen keine Ausprägung fehlt
a_{i-}^{cv} bzw. a_{-k}^{cv}	Objektvektor i bzw. Merkmalsvektor k der Matrix A^{cv}
$A^{mv} = (a_{ik}^{mv})_{n,q}$	Teilmatrix von A jener Merkmale, bei denen mindestens eine Ausprägung fehlt
a_{i-}^{mv} bzw. a_{-k}^{mv}	Objektvektor i bzw. Merkmalsvektor k der Matrix A^{mv}
$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$	Vektor der Kurtosis von m Merkmalen
$B = (b_{ik})_{n,m} = \begin{pmatrix} b_{11} & \dots & b_{1m} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nm} \end{pmatrix}$	Matrix der Hauptkomponenten von A

Symbol	**Bedeutung**
$C = (c_{ij})_{d,r} = \begin{pmatrix} c_{11} & \dots & c_{1r} \\ \vdots & & \vdots \\ c_{d1} & \dots & c_{dr} \end{pmatrix}$	Zuordnungskostenmatrix mit d Spendern und r Empfängern
c_{ij}	Kosten für die Zuordnung von Spender i zu Empfänger j
$\mathfrak{D}$	Diskriminante
$\det(\bullet)$	Determinantenfunktion
d	Anzahl der Spender
D	Spendermenge
dl_i	Donor-Limit spezifisch zum Spender i
dl	Donor-Limit allgemein
dl^{rel}	Donor-Limit relativ zu der Objektmenge
dl^{abs}	Donor-Limit absolut
dl^* bzw. dl^{rel*} bzw. dl^{abs*}	optimales Donor-Limit, allgemein bzw. relativ bzw. absolut
ϵ_j	Offset für Spender j bei der Methode von Siddique und Belin (2008)
$E = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 1 \end{pmatrix}$	Einheitsmatrix
$f(\bullet)$	Dichte- bzw. Wahrscheinlichkeitsfunktion bzw. Verteilung
$\gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix}$	Vektor der Schiefe von m Merkmalen
$g(\bullet)$	Distanzsummenfunktion einer Spender-Empfänger-Zuordnung

Symbol	Bedeutung
h	Index für Objekte
i	Index für Objekte
j	Index für Objekte
k	Anzahl der nächsten Nachbarn bei einer k-Nächste-Nachbarn-Klassifikation
k	Index für Merkmale
l	Index für Merkmale
$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix}$	Diagonalmatrix der Eigenwerte
$\lambda_1, \dots, \lambda_m$	Eigenwerte einer $(m \times m)$ Matrix
$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}$	Vektor der Erwartungswerte von m Merkmalen
M	Menge der Merkmale
M_{ij}	Menge der für Objekte i und j paarweise vorhandenen Merkmale
N	Menge der Objekte
π	Strafterm bei der Methode von Colledge et al. (1978)
p	Skalierungsparameter bei der Minkowski-Distanz
Pr	Wahrscheinlichkeitsmaß
q	Anzahl der Merkmale, für die mindestens eine Ausprägung fehlt

Symbol	**Bedeutung**
$\mathcal{P} = (\rho_{kl})_{m,m} = \begin{pmatrix} \rho_{11} & \dots & \rho_{1m} \\ \vdots & & \vdots \\ \rho_{m1} & \dots & \rho_{mm} \end{pmatrix}$	Korrelationsmatrix
ρ bzw. ρ_{kl}	Korrelation bzw. Korrelation zwischen den Merkmalen k und l
r	Anzahl der Objekte, für die mindestens eine Merkmalsausprägung fehlt, bzw. Empfänger
R	Empfängermenge
$\Sigma = (\sigma_{kl})_{m,m} = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1m} \\ \vdots & & \vdots \\ \sigma_{m1} & \dots & \sigma_{mm} \end{pmatrix}$	Kovarianzmatrix
σ_{kl}	Kovarianz zwischen den Merkmalen k und l
Σ^{cv}	Kovarianzmatrix jener Merkmale, bei denen keine Ausprägung fehlt
$\Sigma^{cv\,-1}$	Inverse der Kovarianzmatrix Σ^{cv}
$(\Sigma^{cv\,-1})_{kl}$	Element aus der k-ten Zeile und der l-ten Spalte von $\Sigma^{cv\,-1}$
$V = (v_{ik})_{n,m} = \begin{pmatrix} v_{11} & \dots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{nm} \end{pmatrix}$	MD-Indikatormatrix
$v_{ik} = \begin{cases} 1 & \text{falls } a_{ik} \text{ fehlt} \\ 0 & \text{sonst} \end{cases}$	MD-Indikator der Ausprägung des Merkmals k bei Objekt i
$V^{cc} = (v_{ik})_{n-r,m}$	Teilmatrix von V jener Objekte, bei denen keine Merkmalsausprägung fehlt
$V^{mc} = (v_{ik})_{r,m}$	Teilmatrix von V jener Objekte, bei denen mindestens eine Merkmalsausprägung fehlt

Symbol	**Bedeutung**
$V^{cv} = (v_{ik})_{n,m-q}$	Teilmatrix von V jener Merkmale, bei denen keine Ausprägung fehlt
$V^{mv} = (v_{ik})_{n,q}$	Teilmatrix von V jener Merkmale, bei denen mindestens eine Ausprägung fehlt
v^{mis} bzw. $\tilde{v}^{mis}$	Anzahl bzw. Anteil der fehlenden Daten in der Datenmatrix
$v_{i\bullet}^{mis}$ bzw. $\tilde{v}_{i\bullet}^{mis}$	Anzahl bzw. Anteil der fehlenden Daten bei Objekt i
$v_{\bullet k}^{mis}$ bzw. $\tilde{v}_{\bullet k}^{mis}$	Anzahl bzw. Anteil der fehlenden Daten bei Merkmal k
w_i	Stichprobengewicht für Objekt i
$x_{ij} = \begin{cases} 1 & \text{falls } D_i \; R_j \text{ zugeordnet} \\ 0 & \text{sonst} \end{cases}$	Zuordnungsindikator
$X = (x_{ij})_{d,r} = \begin{pmatrix} x_{11} & \dots & x_{1r} \\ \vdots & & \vdots \\ x_{d1} & \dots & x_{dr} \end{pmatrix}$	Zuordnungsmatrix
Φ	Matrix der Eigenvektoren
Φ_{-k}	k-ter Eigenvektor einer $(m \times m)$ Matrix

Symbole in Kapitel 4	**Bedeutung**
d	Cohens Effektstärke für unabhängige Stichproben (Abschnitt 4.1.1.2)
$\Delta\bar{p}_1$ bzw. $\Delta\bar{p}_2$	Mittelwert der Δp für zwei verschiedene Donor-Limits
s_1^2 bzw. s_2^2	Varianzen der Δp für zwei verschiedene Donor-Limits
Δp	$(p_I - p_T)/(p_T)$
p_T	Wahrer Wert eines bestimmten Verteilungsparameters
p_I	Schätzwert von p_T nach Imputation
d'	Cohens Effektstärke für verbundene Stichproben (Abschnitt 4.2.1.3)
$\Delta\bar{p}'$	Mittelwert der $\Delta p'_j$
$s^{2\prime}$	Varianz der $\Delta p'_j$
$\Delta p'_j$	Abweichung zwischen $\lvert\Delta p_j^1\rvert$ und $\lvert\Delta p_j^\infty\rvert$
Δp_j^1 bzw. Δp_j^∞	$p_{Tj}^1 - p_{Ij}^1$ bzw. $p_{Tj}^\infty - p_{Ij}^\infty$
p_{Tj}	Wahrer Wert eines bestimmten Verteilungsparameters in Simulationsiteration j
p_{Ij}^1 bzw. p_{Ij}^∞	Schätzwerte von p_{Tj} nach einer Hot-Deck-Imputation mit einem Donor-Limit von 1 bzw. ∞

Abkürzungsverzeichnis

Abkürzung	Bedeutung
ACS	American Community Survey
APA	American Psychological Association
bzw.	beziehungsweise
CART	Classification and Regression Trees
CHAID	Chi-square Automatic Interaction Detectors
CHMP	Committee for Medicinal Products for Human Use
CPS	Current Population Survey
CPS-ORG	Current Population Survey Outgoing Rotation Group
CSV	comma separated values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EM	expectation maximization
EMA	European Medicines Agency
et al.	et alii
FDA	Food and Drug Administration
f.	folgende
ff.	fortfolgende

Abkürzung	Bedeutung
GEIS	Generalized Edit and Imputation System
ID3	Iterative Dichotomiser 3
LFS	Labour Force Survey
MAR	missing at random
MCAR	missing completely at random
MD	Missing Data
MKQ	Methode der kleinsten Quadrate
MNAR	missing not at random
MODI	modified distribution method
NMAR	not missing at random
NOTRA	normal to anything
RMSE	root mean square error
SHS	Survey of Household Spending
SIPP	Survey of Income and Program Participation
SLID	Survey of Labour and Income Dynamics
SQL	structured query language
S.	Seite(n)
USA	United States of America
U.S.	United States
vgl.	vergleiche
v(s).	versus
z. B.	zum Beispiel

Literaturverzeichnis

Abdel-Halim, R. und Abdel-Aal, R. (1999). „Classification of urinary stones by cluster analysis of ionic composition data“. In: *Computer Methods and Programs in Biomedicine* 58, S. 69–81.

Addison, C., Allwright, J., Binsted, N., Bishop, N., Carpenter, B., Dalloz, P., Gee, D., Getov, V., Hey, T., Hockney, R., Lemke, M., Merlin, J., Pinches, M., Scott, C. und Wolton, I. (1993). „The Genesis distributed-memory benchmarks. Part 1: Methodology and general relativity benchmark with results for the SUPRENUM computer“. In: *Concurrency: Practice and Experience* 5, S. 1–22.

Advanced Micro Devices (2013). *Software Optimization Guide for AMD Family 15h Processors*. Number 47414, Revision 3.06. Advanced Micro Devices, Inc.

Allison, P. (2001). *Missing Data*. Thousand Oaks: Sage.

American Psychological Association (2001). *Publication Manual of the American Psychological Association (5th ed.)*. Washington, D.C.: American Psychological Association.

Andridge, R. und Little, R. (2010). „A review of hot deck imputation for survey non-response“. In: *International Statistical Review* 78, S. 40–64.

Azen, S. und van Guilder, M. (1981). „Conclusions regarding algorithms for handling incomplete data“. In: *Proceedings of the Statistical Computing Section*. The American Statistical Association, S. 53–56.

Backhaus, K. und Blechschmidt, B. (2009). „Fehlende Werte und Datenqualität“. In: *Die Betriebswirtschaft* 69, S. 265–287.

Bankhofer, U. (1995). *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse*. Köln: Eul.

Bankhofer, U. und Joenssen, D. (2014). „On limiting donor usage for imputation of missing data via hot deck methods“. In: *Data Analysis, Machine*

Learning and Knowledge Discovery. Hrsg. von M. Spiliopoulou, L. Schmidt-Thieme und R. Jannings. Berlin: Springer, S. 3–11.

Barak, A. und Fried, C. (2002). „Leading cases, I. constitutional law, A. census clause“. In: *Harvard Law Review* 116, S. 200–210.

Barzi, F. und Woodward, M. (2004). „Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies“. In: *American Journal of Epidemiology* 160, S. 34–45.

Beaumont, J. und Bissonnette, J. (2011). „Variance estimation under composite imputation: The methodology behind SEVANI“. In: *Survey Methodology* 37, S. 171–179.

Bodner, T. (2006). „Missing data: Prevalence and reporting practices“. In: *Psychological Reports* 99, S. 675–680.

Bollinger, C. und Hirsch, B. (2006). „Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching“. In: *Journal of Labor Economics* 24, S. 483–519.

Borz, J. und Döring, N. (2009). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer.

Breiman, L., Friedman, J., Stone, C. und Olshen, R. (1984). *Classification and Regression Trees*. Monterey: Wadsworth und Brooks.

Brick, J. und Kalton, G. (1996). „Handling missing data in survey research“. In: *Statistical Methods in Medical Research* 5, S. 215–238.

Brittingham, A. (1998). *National Household Survey on Drug Abuse: Main Findings, 1996*. Darby: DIANE Publishing Company.

Buck, S. (1960). „A method of estimation of missing values in multivariate data suitable for use with an electronic computer“. In: *Journal of the Royal Statistical Society Series B (Methodological)* 22, S. 302–306.

Cantwell, P., Hogan, H. und Styles, K. (2004). „The use of statistical methods in the U.S. Census“. In: *The American Statistician* 58, S. 203–212.

Cario, M. und Nelson, B. (1997). „Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix“. In: *Northwestern University, IEMS Technical Reports* 50, S. 100–150.

Chauvet, G., Deville, J. und Haziza, D. (2011). „On balanced random imputation in surveys“. In: *Biometrika* 98, S. 459–471.

Chen, J. und Shao, J. (2001). „Jackknife variance estimation for nearest-neighbor imputation“. In: *Journal of the American Statistical Association* 96, S. 260–269.

Chen, J., Rao, J. und Sitter, R. (2000). „Efficient random imputation for missing data in complex surveys“. In: *Statistica Sinica* 10, S. 1153–1170.

Codd, E. (1979). „Extending the database relational model to capture more meaning“. In: *ACM Transactions on Database Systems* 4, S. 397–434.

Coder, J. (1978). „Income data collection and processing from the march income supplement to the Current Population Survey“. In: *The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing*. Hrsg. von D. Kasprzyk. Washington, D.C.: Department of Health, Education und Welfare.

Cohen, J. (1997). „A power primer“. In: *Quantitative Methods in Psychology* 112, S. 155–159.

Colledge, M., Johnson, J. und Sande, I. (1978). „Large scale imputation of survey data“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 431–436.

Collins, L., Schafer, J. und Kam, C. (2001). „A comparison of inclusive and restrictive strategies in modern missing data procedures“. In: *Psychological Methods* 6, S. 330–351.

Committee for Medicinal Products for Human Use (2010). *Guideline on Missing Data in Confirmatory Clinical Trials*. London: European Medicine Agency.

Coutinho, W. und de Waal, T. (2012). *Hot deck imputation of numerical data under edit restrictions*. Discussion paper. The Hague: Statistics Netherlands.

Cox, B. (1980). „The weighted sequential hot deck imputation procedure“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 721–726.

Creel, D. und Krotki, K. (2006). „Creating imputation classes using classification tree methodology“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 2884–2887.

Cresce Jr., A., Obenski, S. und Chappell, G. (2005). „Research to improve census imputation methods: The plan to examine count and item imputati-

on“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 2928–2934.

Dahl, F. (2007). „Convergence of random-nearest-neighbour imputation“. In: *Computational Statistics & Data Analysis* 51, S. 5913–5917.

de Waal, T., Pannekoek, J. und Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken: Wiley.

Dear, R. (1959). *A principal component missing data method for multiple regression models*. Technical Report SP-86. System Development Corporation.

Dempster, A., Laird, N. und Rubin, D. (1977). „Maximum likelihood from incomplete data via the EM algorithm“. In: *Journal of the Royal Statistical Society Series B (Methodological)* 39, S. 1–38.

Deza, E. und Deza, M. (2006). *Dictionary of Distances*. Amsterdam: Elsevier.

Dickson, P. und Allen, T. (2006). *The Bonus Army: An American Epic*. New York: Walker.

Dillman, D., Eltinge, J., Groves, R. und Little, R. (2002). „Survey nonresponse in design, data collection, and analysis“. In: *Survey Nonresponse*. Hrsg. von R. Groves, D. Dillman, J. Eltinge und R. Little. New York: John Wiley und Sons, S. 3–26.

Domschke, W. (1995). *Logistik: Transport*. München: Oldenbourg.

Durrant, G. (2009). „Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates“. In: *International Journal of Social Research Methodology* 12, S. 293–304.

Durrant, G. und Skinner, C. (2006). „Using missing data methods to correct for measurement error in a distribution function“. In: *Survey Methodology* 32, S. 25–36.

Enders, C. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press.

Ester, M., Kriegel, H., Sander, J. und Xu, X. (1996). „A density-based algorithm for discovering clusters in large spatial databases with noise“. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, S. 226–231.

Fahrmeir, L., Hamerle, A. und Tutz, G. (1996a). *Multivariate Statistische Verfahren*. Berlin: de Gruyter.

Fahrmeir, L., Kaufmann, H. und Kredler, C. (1996b). „Regressionsanalyse“. In: *Multivariate Statistische Verfahren*. Hrsg. von L. Fahrmeir, A. Hamerle und G. Tutz. Berlin: de Gruyter, S. 93–168.

Fay, R. (1999). „Theory and application of nearest neighbor imputation in Census 2000“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 112–121.

Fix, E. und Hodges, J. (1952). *Discriminatory Analysis: Nonparametric Discrimination: Small Sample Performance*. Technical Report: Project 21-49-004, Report Number 11. Randolph Field, Texas: USAF School of Aviation Medicine.

Ford, B. (1983). „An overview of hot-deck procedures“. In: *Incomplete Data in Sample Surveys*. Hrsg. von W. Madow, H. Nisselson und I. Olkin. Volume 2. New York: Academic Press, S. 185–207.

Fröhlich, M. und Pieter, A. (2009). „Cohen's Effektstärken als Mass der Bewertung von praktischer Relevanz – Implikationen für die Praxis“. In: *Schweizerische Zeitschrift für Sportmedizin und Sporttraumatologie* 57, S. 139–142.

Gelman, A. und Rubin, D. (1992). „Inference from iterative simulation using multiple sequences“. In: *Statistical Science* 7, S. 457–472.

Geman, S. und Geman, D. (1984). „Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, S. 721–741.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. und Hothorn, T. (2013). *mvtnorm: Multivariate Normal and t Distributions*. R-Package Version 0.9-9995. URL: http://CRAN.R-project.org/package=mvtnorm.

Gini, C. (1912). „Variabilità e mutabilità“. In: *Memorie di Metodologica Statistica (Ed. Pizetti E, Salvemini, T). Rom: Libreria Eredi Virgilio Veschi* 1, (Neudruck in 1955).

Glover, F., Karney, D., Klingman, D. und Napier, A. (1974). „A computation study on start procedures, basis change criteria, and solution algorithms for transportation problems“. In: *Management Science* 20, S. 793–813.

Glynn, R., Laird, N. und Rubin, D. (1993). „Multiple imputation in mixture models for nonignorable nonresponse with follow-ups“. In: *Journal of the American Statistical Association* 88, S. 984–993.

Graham, J. (2009). „Missing data analysis: Making it work in the real world“. In: *Annual Review of Psychology* 60, S. 549–576.

Graham, J. (2012). *Missing Data: Analysis and Design*. New York: Springer.

Graham, J. und Donaldson, S. (1993). „Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data“. In: *Journal of Applied Psychology* 78, S. 119–128.

Graham, J., Taylor, B., Olchowski, A. und Cumsille, P. (2006). „Planned missing data designs in psychological research“. In: *Psychological Methods* 11, S. 323–343.

Grau, E. und Ahmed, S. (2008). „Evaluation of imputation of covariates in an impact analysis with regression adjustment“. In: *Proceedings of the Section on Health Policy Statistics*. The American Statistical Association, S. 114–124.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. und Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley und Sons.

Haitovsky, Y. (1968). „Missing data in regression analysis“. In: *Journal of the Royal Statistical Society Series B (Methodological)* 30, S. 67–81.

Hamaker, H. und van Strik, R. (1955). „The efficiency of double dampling for attributes“. In: *Journal of the American Statistical Association* 50, S. 830–849.

Haziza, D. (2007). „Variance estimation for a ratio in the presence of imputed data“. In: *Survey Methodology* 33, S. 159–166.

Herzog, T., Scheuren, F. und Winkler, W. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Hogye, M. (2004). „Searching for donors: Finding an imputation strategy“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 3669–3676.

Huckett, J. und Larsen, M. (2007). „Microdata simulation for confidentiality protection using regression quantiles and hot deck“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 3053–3060.

Huisman, M. (2000). „Imputation of missing item responses: Some simple techniques“. In: *Quality & Quantity* 34, S. 331–351.

Janes, D. (2007). „Can a geographic sort improve hot deck donor imputation in the Canadian Census?“ In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 1479–1485.

Jhun, M., Jeong, H. und Koo, J. (2007). „On the use of adaptive nearest neighbors for missing value imputation“. In: *Communications in Statistics: Simulation and Computation* 36, S. 1275–1286.

Jobson, J. (1991). *Applied Multivariate Data Analysis: Regression and Experimental Design*. Volume 1. New York: Springer.

Jobson, J. (1992). *Applied Multivariate Data Analysis: Categorical and Multivariate Methods*. Volume 2. New York: Springer.

Joenssen, D. (2013). *HotDeckImputation: Hot Deck Imputation Methods for Missing Data*. R-Package Version 0.1.0. URL: http://CRAN.R-project.org/package=HotDeckImputation.

Joenssen, D. und Bankhofer, U. (2012). „Donor limited hot deck imputation: Effects on parameter estimation“. In: *Journal of Theoretical and Applied Computer Science* 6, S. 58–70.

Joenssen, D. und Müllerleile, T. (2014). „Fehlende Daten beim Data-Mining“. In: *HMD Praxis der Wirtschaftsinformatik* 51, S. 458–468.

Jönsson, P. und Wohlin, C. (2006). „Benchmarking k-nearest neighbour imputation with homogeneous Likert data“. In: *Empirical Software Engineering* 11, S. 463–489.

Judkins, D. (1997). „Imputing for Swiss cheese patterns of missing data“. In: *Proceedings of the Symposium: New Directions in Surveys and Censuses*. Statistics Canada, S. 143–148.

Kaiser, J. (1983). „The effectiveness of hot-deck procedures in small samples“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 523–528.

Kaiser, J. (1989). „The robustness of hot-deck and cell mean methods in retaining population covariance structure in imputed samples“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 286–289.

Kalton, G. und Kasprzyk, D. (1982). „Imputing for missing survey responses“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 22–31.

Kalton, G. und Kasprzyk, D. (1986). „The treatment of missing survey data“. In: *Survey Methodology* 12, S. 1–16.

Kalton, G. und Kish, L. (1981). „Two efficient random imputation procedures“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 146–151.

Kalton, G. und Kish, L. (1984). „Some efficient random imputation methods“. In: *Communications in Statistics: Theory and Methods* 13, S. 1919–1939.

Kass, G. (1980). „An exploratory technique for investigating large quantities of categorical data“. In: *Applied Statistics* 29, S. 119–127.

Kim, J. und Curry, J. (1977). „The treatment of missing data in multivariate analysis“. In: *Sociological Methods and Research* 6, S. 215–240.

Kim, J. und Fuller, W. (2004). „Fractional hot deck imputation“. In: *Biometrika* 91, S. 559–578.

Koenker, R. und Bassett, G. (1978). „Regression quantiles“. In: *Econometrica* 46, S. 33–50.

Kovar, J. und Whitridge, J. (1995). „Imputation of business survey data“. In: *Business Survey Methods*. Hrsg. von B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge und P. Kott. New York: Wiley, S. 403–423.

Kozak, R. (2005). „The Banff system for automated editing and imputation“. In: *Proceedings of the Survey Research Methods Section*. Statistical Society of Canada, S. 1–10.

Little, R. (1988). „Missing-data adjustments in large surveys“. In: *Journal of Business and Economic Statistics* 6, S. 287–296.

Little, R. (1992). „Regression with missing X's: A review“. In: *Journal of the American Statistical Association* 87, S. 1227–1237.

Little, R. und Rubin, D. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: John Wiley und Sons.

Little, R. (1982). „Models for nonresponse in sample surveys“. In: *Journal of the American Statistical Association* 77, S. 237–250.

Longford, N. (2005). *Missing Data and Small-Area Estimation*. New York: Springer.

Mahalanobis, P. (1930). „On tests and measures of group divergence“. In: *Journal of the Asiatic Society of Bengal* 39, S. 541–588.

Mahalanobis, P. (1936). „On the generalised distance in statistics“. In: *Proceedings of the National Institute of Sciences of India* 1, S. 49–55.

Marker, D., Judkins, D. und Winglee, M. (2002). „Large-scale imputation for complex surveys“. In: *Survey Nonresponse*. Hrsg. von R. Groves, D. Dillman, J. Eltinge und R. Little. New York: Wiley, S. 329–341.

Matsumoto, M. und Nishimura, T. (1998). „Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator“. In: *ACM Transactions on Modeling and Computer Simulation* 8, S. 3–30.

Matthai, A. (1951). „Estimation of parameters from incomplete data with application to design of sample surveys“. In: *Sankhya* 2, S. 145–152.

McKnight, P., McKnight, K., Sidani, S. und Figueredo, A. (2007). *Missing Data: A Gentle Introduction*. New York: Guilford Press.

Müllerleile, T. und Nissen, V. (2014). „When processes alienate customers: Towards a theory of process acceptance“. In: *S-BPM ONE - Scientific Research*. Hrsg. von A. Nanopoulos und W. Schmidt. Berlin: Springer, S. 171–180.

Nordholt, E. (1998). „Imputation: Methods, simulation experiments and practical examples“. In: *International Statistical Review* 66, S. 157–180.

Oh, H. und Scheuren, F. (1983). „Weighting adjustment for unit nonresponse“. In: *Incomplete Data in Sample Surveys*. Hrsg. von W. Madow, H. Nisselson und I. Olkin. Volume 2. New York: Academic Press, S. 143–184.

Oh, H. und Scheuren, F. (1980). „Estimating the variance impact of missing CPS income data“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 408–415.

O'Kelly, M. und Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners*. Statistics in Practice. New York: Wiley.

Ono, M. und Miller, H. (1969). „Income nonresponses in the current population survey“. In: *Proceedings of the Social Statistics Section*. The American Statistical Association, S. 277–288.

Panel on Handling Missing Data in Clinical Trials (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, D.C.: The National Academies Press.

Pearson, K. (1900). „On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that

it can be reasonably supposed to have arisen from random sampling“. In: *Philosophical Magazine* 50, S. 157–175.

Peláez, J., Doña, J. und La Red, D. (2008). „Fuzzy imputation method for database systems“. In: Hrsg. von J. Peláez, J. Doña und D. La Red. Hershey: IGI Global, S. 805–821.

Peughd, J. und Enders, C. (2004). „Missing data in educational research: A review of reporting practices and suggestions for improvement“. In: *Review of Educational Research* 74, S. 525–556.

Pokropp, F. (1996). *Stichproben: Theorie und Verfahren*. München: Oldenbourg.

Popper, K. (1968). *The Logic of Scientific Discovery*. New York: Harper Torchbooks.

Quinlan, J. (1979). „Discovering rules by induction from large collections of examples“. In: *Expert Systems in the Micro Electronic Age*. Hrsg. von D. Michie. Edinburgh: Edinburgh University Press, S. 168–201.

Quinlan, J. (1983). „Learning efficient classification procedures and their application to chess endgames“. In: *Machine Learning: An Artificial Intelligence Approach*. Hrsg. von R. Michalski, J. Carbonell und T. Mitchell. Palo Alto: Tioga Publishing Company, S. 463–482.

Quinlan, J. (1986). „Induction of decision trees“. In: *Machine Learning* 1, S. 81–106.

Quinlan, J. (1993). *C4.5. Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raghunathan, T. und Grizzle, J. (1995). „A split questionnaire survey design“. In: *Journal of the American Statistical Association* 90, S. 54–63.

Rancourt, E. (1999). „Estimation with nearest neighbour imputation at Statistics Canada“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 131–138.

Rao, J. und Shao, J. (1992). „Jackknife variance estimation with survey data under hot deck imputation“. In: *Biometrika* 79, S. 811–822.

Reinfeld, N. und Vogel, W. (1958). *Mathematical Programming*. New Jersey: Prentice-Hall.

Rizvi, M. (1983). „Hot-deck procedures: Introduction“. In: *Incomplete Data in Sample Surveys*. Hrsg. von W. Madow, H. Nisselson und I. Olkin. Volume 3. New York: Academic Press, S. 351–352.

Rosenbaum, P. und Rubin, D. (1983). „The central role of the propensity score in observational studies for causal effects“. In: *Biometrika* 70, S. 41–55.

Rosenbaum, P. und Rubin, D. (1985). „Constructing a control group using multivariate matched sampling methods that incorporate the propensity score“. In: *The American Statistician* 39, S. 33–38.

Roth, P., Switzer, F. und Switzer, D. (1999). „Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques“. In: *Organizational Research Methods* 2, S. 211–232.

Roth, P. (1994). „Missing data: A conceptual review for applied psychologists“. In: *Personnel Psychology* 47, S. 537–560.

Roth, P. und Switzer III, F. (1995). „A Monte Carlo analysis of missing data techniques in a HRM setting“. In: *Journal of Management* 21, S. 1003–1023.

Rubin, D. (1976a). „Inference and missing data“. In: *Biometrika* 63, S. 581–592.

Rubin, D. (1976b). „Multivariate matching methods that are equal percent bias reducing, I: Some examples“. In: *Biometrics* 32, S. 109–120.

Rubin, D. (1977). *The design of a general and flexible system for handling non-response in sample surveys*. Working Paper. U.S. Social Security Administration.

Rubin, D. (1978). „Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonrepsonse“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 20–34.

Rubin, D. (1979). „Using multivariate matched sampling and regression adjustment to control bias in observational studies“. In: *Journal of the American Statistical Association* 74, S. 318–328.

Rubin, D. (1980). „Bias reduction using Mahalanobis' metric matching“. In: *Biometrics* 36, S. 295–298.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. (1996). „Multiple imputation after 18+ years“. In: *Journal of the American Statistical Association* 91, S. 473–489.

Rubin, D. und Thomas, N. (1992). „Characterizing the effect of matching using linear propensity score methods with normal distributions“. In: *Biometrika* 79, S. 797–809.

Rüger, B. (2002). *Statistische Tests*. München: Oldenbourg.

Ruggles, S., Alexander, J., Genadek, K., Goeken, R., Schroeder, M. und Sobek, M. (2010). *Integrated public use microdata series: Version 5.0*. Machine-readable database. Minneapolis: Minnesota Population Center.

Sande, I. (1983). „Hot-deck imputation procedures“. In: *Incomplete Data in Sample Surveys*. Hrsg. von W. Madow, H. Nisselson und I. Olkin. Volume 3. New York: Academic Press, S. 339–349.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman und Hall.

Schafer, J. und Graham, J. (2002). „Missing data: Our view of the state of the art“. In: *Psychological Methods* 7, S. 147–177.

Schenker, N. und Taylor, J. (1996). „Partially parametric techniques for multiple imputation“. In: *Computational Statistics & Data Analysis* 22, S. 425–446.

Schlomer, G., Bauman, S. und Card, N. (2010). „Best practicies for missing data management in counseling psychology“. In: *Journal of Counseling Psychology* 57, S. 1–10.

Schnell, R. (1986). *Missing-Data Probleme in der Empirischen Sozialforschung*. Bochum: Dissertation.

Schwab, G. (1991). *Fehlende Werte in der Angewandten Statistik*. Wiesbaden: Deutscher Universitäts-Verlag.

Shannon, C. und Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shao, J. (2000). „Cold deck and ratio imputation“. In: *Survey Methodology* 26, S. 79–85.

Shao, J. und Sitter, R. (1996). „Bootstrap for imputed survey data“. In: *Journal of the American Statistical Association* 91, S. 1278–1288.

Siddique, J. und Belin, T. (2008). „Multiple imputation using an iterative hot-deck with distance-based donor selection“. In: *Statistics in Medicine* 27, S. 83–102.

Spiess, M. (2005). „Analyse von Längsschnittdaten mit fehlenden Werten: Grundlagen, Verfahren und Anwendungen“. Habilitation. Fachbereich 8: Sozialwissenschaften der Universität Bremen.

Srinivasan, V. und Thompson, G. (1973). „Benefit-cost analysis of coding techniques for the primal transportation algorithm“. In: *Journal of the ACM* 20, S. 194–213.

Statistical Research Group (1948). *Sampling Inspection*. New York: McGraw-Hill.

Strassen, V. (1969). „Gaussian elimination is not optimal“. In: *Numerische Mathematik* 13, S. 354–356.

Strike, K., Emam, K. und Madhavji, N. (2001). „Software cost estimation with incomplete data“. In: *IEEE Transactions on Software Engineering* 27, S. 890–908.

Sydow, J. (1985). *Der Soziotechnische Ansatz der Arbeits- und Organisationsgestaltung: Darstellung, Kritik, Weiterentwicklung*. Frankfurt am Main: Campus Verlag.

Tang, G., Little, R. und Raghunathan, T. (2003). „Analysis of multivariate missing data with nonignorable nonresponse“. In: *Biometrika* 90, S. 747–764.

Thran, S. und Gillis, K. (1992). „A comparison of imputation techniques in a physician survey“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 211–222.

Trefethen, L. und Bau, D. (1997). *Numerical Linear Algebra*. Philadelphia: SIAM.

Trist, E. und Bamford, K. (1951). „Some social and psychological consequences of the longwall method of coal-getting“. In: *Human Relations* 4, S. 3–38.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. und Altman, R. (2001). „Missing value estimation methods for DNA microarrays“. In: *Bioinformatics* 17, S. 520–525.

Ueberhuber, C. (1997). *Numerical Computation 1: Methods, Software, and Analysis*. Berlin: Springer.

U.S. Bureau of the Census (2006). *Current Population Survey: Design and Methodology, Technical Paper 66*. Washington D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census (2010). *2009 Data Release, Data & Documentation. American Community Survey, U.S. Census Bureau*. URL: https://www.census.gov/acs/www/data_documentation/2009_release/.

Wacholder, S., Carroll, R., Pee, D. und Gail, M. (1994). „The partial questionnaire design for casecontrol studies (with discussion)“. In: *Statistics in Medicine* 13, 623–649.

West, S., Butani, S. und Witt, M. (1990). „Alternative imputation methods for wage data“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 254–259.

Whitridge, P. und Kovar, J. (1990). „Applications of the generalized edit and imputation system at Statistics Canada“. In: *Proceedings of the Survey Research Methods Section*. The American Statistical Association, S. 105–110.

Wilkinson, L. und Task Force on Statistical Inference (1999). „Statistical methods in psychology journals: Guidelines and explanations“. In: *American Psychologist* 54, S. 594–604.

Wilks, S. (1932). „Moments and distributions of estimates of population parameters from fragmentary samples“. In: *The Annals of Mathematical Statistics* 3, S. 163–195.

Wood, A., White, I. und Thompson, S. (2004). „Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals“. In: *Clinical Trials* 1, S. 368–376.

Yenduri, S. und Iyengar, S. (2007). „Performance evaluation of imputation methods for incomplete datasets“. In: *International Journal of Software Engineering and Knowledge Engineering* 17, S. 127–152.